

# **Computational Methods for Interactive and Explorative Study Design and Integration of High-throughput Biological Data**

**Dissertation**

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

vorgelegt von

**M. Sc. Andreas Friedrich**

aus Erfurt

Tübingen

2021

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

Tag der mündlichen Qualifikation:

Dekan:

1. Berichterstatter:

2. Berichterstatterin:

17.12.2021

Prof. Dr. Thilo Stehle

Prof. Dr. Oliver Kohlbacher

Prof. Dr. Kay Nieselt

*“Possibly, but my concern is that there not be more things in my  
philosophy than are in heaven and earth.”*

- Willard van Orman Quine (1908 – 2000)

in response to being quoted William Shakespeare’s statement from Hamlet:

"There are more things in heaven and earth... than are dreamt of in your philosophy."



# Erklärung

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

*Computational Methods for Interactive and Explorative Study Design and Integration of High-throughput Biological Data*

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.



# Abstract

The increase in the use of high-throughput methods to gain insights into biological systems has come with new challenges. Genomics, transcriptomics, proteomics, and metabolomics lead to a massive amount of data and metadata. While this wealth of information has resulted in many scientific discoveries, new strategies are needed to cope with the ever-growing variety and volume of metadata. Despite efforts to standardize the collection of study metadata, many experiments cannot be reproduced or replicated. One reason for this is the difficulty to provide the necessary metadata. The large sample sizes that modern omics experiments enable, also make it increasingly complicated for scientists to keep track of every sample and the needed annotations. The many data transformations that are often needed to normalize and analyze omics data require a further collection of all parameters and tools involved. A second possible cause is missing knowledge about statistical design of studies, both related to study factors as well as the required sample size to make significant discoveries.

In this thesis, we develop a multi-tier model for experimental design and a portlet for interactive web-based study design. Through the input of experimental factors and the number of replicates, users can easily create large, factorial experimental designs. Changes or additional metadata can be quickly uploaded via user-defined spreadsheets including sample identifiers. In order to comply with existing standards and provide users with a quick way to import existing studies, we provide full interoperability with the ISA-Tab format. We show that both data model and portlet are easily extensible to create additional tiers of samples annotated with technology-specific metadata.

We tackle the problem of unwieldy experimental designs by creating an aggregation graph. Based on our multi-tier experimental design model, similar samples, their sources, and analytes are summarized, creating an interactive summary graph that focuses on study factors and replicates. Thus, we give researchers a quick overview of sample sizes and the aim of different studies. This graph can be included in our portlets or used as a stand alone application and is compatible with the ISA-Tab format. We show that this approach can be used to explore the quality of publicly available experimental designs and metadata annotation.

The third part of this thesis contributes to a more statistically sound experiment planning for differential gene expression experiments. We integrate two tools for the prediction of statistical power and sample size estimation into our portal. This integra-

---

tion enables the use of existing data, in order to arrive at more accurate calculation for sample variability. Additionally, the statistical power of existing experimental designs of certain sample sizes can be analyzed. All results and parameters are stored and can be used for later comparison.

Even perfectly planned and annotated experiments cannot eliminate human error. Based on our model we develop an automated workflow for microarray quality control, enabling users to inspect the quality of normalization and cluster samples by study factor levels. We import a publicly available microarray dataset to assess our contributions to reproducibility and explore alternative analysis methods based on statistical power analysis.

# Kurzfassung

Die verstärkte Nutzung von Hochdurchsatz-Methoden, um Erkenntnisse über biologische Systeme zu gewinnen, hat zu neuen Herausforderung geführt. Genomik, Transkriptomik, Proteomik und Metabolomik erzeugen gewaltige Mengen an Daten und Metadaten. Während dieser Datenreichtum zu vielen wissenschaftlichen Entdeckungen geführt hat, werden neue Strategien benötigt, um die stetig wachsende Vielfalt und das Volumen an Metadaten zu bewältigen.

Trotz Anstrengungen die Sammlung von Metadaten zu standardisieren, können viele Experimente nicht reproduziert oder repliziert werden. Eine Ursache hierfür ist die Schwierigkeit die nötigen Metadaten bereitzustellen. Die große Anzahl an Proben, die moderne Omics-Experimente ermöglichen, macht es gleichzeitig immer schwieriger für Wissenschaftler die Übersicht über jede einzelne Probe und die nötige Annotation zu behalten. Die vielen Daten-Transformationen, die oft nötig sind, um Omics-Daten zu normalisieren und analysieren, machen die Erfassung aller verwendeten Parameter und Tools nötig. Eine zweite mögliche Ursache ist fehlendes Wissen über statistisches Design von Studien, sowohl was Studien-Faktoren als auch die für aussagekräftige Entdeckungen benötigte Stichprobengröße betrifft.

In dieser Arbeit entwickeln wir ein mehrschichtiges Modell für experimentelles Design und ein Portlet für interaktives web-basiertes Studiendesign. Durch die Eingabe von experimentellen Faktoren und der Zahl von Replikaten können Nutzer leicht große, faktorielle Experimental-Designs erstellen. Änderungen oder zusätzliche Metadaten können schnell über ein von Nutzern definiertes Tabellen-Format hochgeladen werden, welches Proben-Identifikatoren enthält. Um die Nutzung existierender Standards zu gewährleisten und Nutzern eine schnelle Möglichkeit zum Import existierender Studien zu geben, stellt unser Ansatz komplette Interoperabilität mit dem ISA-Tab-Format bereit. Wir zeigen, dass sowohl unser Datenmodell, als auch unser Portlet leicht erweiterbar sind, um zusätzliche Level aus Proben zu beschreiben, die mit technologie-spezifischen Metadaten annotiert sind.

Wir überkommen das Problem von Experimental-Designs hinderlicher Größe, indem wir einen Aggregations-Graphen entwickeln. Basierend auf unserem mehrschichtigen Modell für experimentelles Designs, fassen wir ähnliche Proben, deren Ursprungs-Organismen, und die gemessenen Analyte zusammen, um einen interaktiven Graph zu erzeugen, dessen Fokus auf experimentellen Faktoren der Studie und auf der Zahl

---

von Replikaten liegt. Damit geben wir Forschern einen schnellen Überblick über die Stichprobengröße und das Ziel einer Studie. Der Graph kann in unsere Portlets integriert oder als eigenständige Anwendung genutzt werden und ist kompatibel mit dem ISA-Tab-Format. Wir zeigen, dass unser Ansatz genutzt werden kann, um öffentlich abrufbare experimentelle Designs auf ihre Qualität und die Vollständigkeit ihrer Annotation zu untersuchen.

Der dritte Teil dieser Arbeit leistet einen Beitrag zu einer statistisch standfesten Planung von Experimenten zur differentiellen Analyse von Genexpression. Wir integrieren zwei Tools für die Vorhersage von Teststärke und für die Abschätzung der benötigten Stichprobengröße in unser Portal. Diese Integration ermöglicht es bestehende Daten zu nutzen, um akkuratere Berechnungen für Proben-Variabilität zu erhalten. Zusätzlich kann die Teststärke bestehender experimenteller Designs mit bestimmten Probengrößen analysiert werden. Alle Resultate und Parameter werden gespeichert und können für spätere Vergleiche genutzt werden.

Sogar perfekt geplante und annotierte Experimente können menschliche Fehler nicht beseitigen. Basierend auf unserem Modell entwickeln wir einen automatisierten Workflow für die Qualitätskontrolle von Microarray-Messungen. Wir ermöglichen Nutzern die Qualität der Normalisierung zu untersuchen und Proben nach den Levels der Studien-Faktoren zu clustern. Wir importieren eine öffentlich zugängliche Microarray-Studie, um unseren Beitrag zur Reproduzierbarkeit zu bewerten und erforschen alternative Analyse-Methoden basierend auf unserer Analyse der Teststärke.

---

Für meine Eltern



# Acknowledgments

I would like to express my sincere gratitude to my advisor Prof. Oliver Kohlbacher for his continuous support, encouragement and guidance in my research and for his feedback during the writing of this thesis. I am also grateful to Prof. Kay Nieselt for her invaluable guidance, especially on the topics of transcriptomics and statistics. I thank both of them for granting me the opportunity to pursue the subjects of this thesis with considerable freedom. Last but not least, I am grateful to Prof. Sven Nahnsen for readily sharing his knowledge on proteomics and for the opportunity to help build something new at the Quantitative Biology Center.

In this respect I also thank Christopher Mohr, David Wojnar, Matthias Seybold and everyone else who helped tame the beast that is openBIS. I also want to thank the openBIS team at the ETH Zürich for their fast and always helpful replies to any of our questions.

I thank all of my former and current colleagues at the Applied Bioinformatics group, the QBiC, and the Integrative Transcriptomics group for interesting discussions, sharing their wisdom and especially for the fun and welcoming working environment, social events, sushi, BBQs and Glühweinparties.

Special thanks go out to everyone reviewing and proofreading this thesis, especially to Jörg Peter and Fabian Aicheler for their extensive and knowledgeable feedback.

Last but not least, I want to thank Luis de la Garza, despite him forgetting the shrimps.

In accordance with the standard scientific protocol, I will use the personal pronoun *we* to indicate the reader and the writer or my scientific collaborators and myself.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Contributions of this Thesis . . . . .	4
<b>2</b>	<b>Background</b>	<b>7</b>
2.1	Big Data and High-Throughput Omics Experiments . . . . .	7
2.1.1	Methods for Differential Gene Expression Analysis . . . . .	10
2.2	Experimental Design . . . . .	14
2.2.1	Factorial Design . . . . .	15
2.2.2	Statistical Power . . . . .	16
2.2.3	Statistical Power of Microarray experiments . . . . .	18
2.2.4	Statistical Power of RNA-Seq experiments . . . . .	20
2.3	Reproducibility and Replicability . . . . .	22
2.3.1	Standards for Experiment Reporting . . . . .	23
2.3.2	Science Gateways and Workflows . . . . .	25
2.3.3	Virtual Machines and Containers . . . . .	26
<b>3</b>	<b>Modeling of Experimental Designs</b>	<b>29</b>
3.1	Introduction . . . . .	29
3.2	Requirements . . . . .	30
3.2.1	Software Interfaces . . . . .	31
3.2.2	Definitions and Terminology . . . . .	32
3.2.3	User Characteristics and User Interface . . . . .	33
3.2.4	Functional Requirements . . . . .	33
3.2.5	Performance Requirements . . . . .	37
3.2.6	Software System Attributes . . . . .	38
3.3	Design and Implementation . . . . .	39
3.3.1	Backend and Data Model . . . . .	39
3.3.2	ISA-Tab Support . . . . .	41
3.3.3	Software Interfaces and User Management . . . . .	43

## Table of Contents

---

3.3.4	Guided Metadata Management . . . . .	43
3.3.5	Software Components . . . . .	44
3.3.6	Data Transfer . . . . .	46
3.4	Results . . . . .	46
3.4.1	Interactive Study Design . . . . .	46
3.4.2	Response Time . . . . .	51
3.4.3	Metadata Management . . . . .	52
3.4.4	Study Import and Interoperability . . . . .	54
3.4.5	Extensibility . . . . .	55
3.5	Discussion . . . . .	57
<b>4</b>	<b>Visual Exploration of Experimental Designs</b>	<b>61</b>
4.1	Introduction . . . . .	61
4.2	Materials and Methods . . . . .	62
4.2.1	Generation of Aggregation Graphs . . . . .	62
4.2.2	Validation on Studies . . . . .	64
4.3	Design and Implementation . . . . .	64
4.4	Results . . . . .	65
4.4.1	qPortal Integration . . . . .	65
4.4.2	Runtime . . . . .	67
4.4.3	Validation . . . . .	68
4.5	Discussion . . . . .	69
<b>5</b>	<b>Interactive Sample Size Calculation for Differential Gene Expression Experiments</b>	<b>73</b>
5.1	Introduction . . . . .	73
5.2	Material and Methods . . . . .	74
5.2.1	RNA-Seq Variance Model . . . . .	74
5.2.2	Parameter Optimization and Sample Size Calculation . . . . .	75
5.3	Implementation . . . . .	75
5.3.1	Backend . . . . .	76
5.4	Results . . . . .	77
5.4.1	Graphical User Interface for Power Analysis of Experimental Designs . . .	77
5.4.2	Validation . . . . .	79
5.5	Discussion . . . . .	81
<b>6</b>	<b>Using Experimental Design for Data Processing and Visualization</b>	<b>85</b>
6.1	Introduction . . . . .	85
6.2	Materials and Methods . . . . .	86
6.2.1	Quality Control Workflow . . . . .	86

---

6.2.2	Differential Expression Analysis . . . . .	86
6.3	Results . . . . .	87
6.3.1	Study Import . . . . .	87
6.3.2	Sample Size Estimation . . . . .	88
6.3.3	Quality Control . . . . .	89
6.3.4	Differential Expression Analysis . . . . .	91
6.4	Discussion . . . . .	94
<b>7</b>	<b>Conclusion and Outlook</b>	<b>97</b>
	<b>Bibliography</b>	<b>101</b>
	<b>Appendices</b>	<b>119</b>
	<b>Appendix A Abbreviations</b>	<b>121</b>
	<b>Appendix B Visual Exploration of Experimental Designs</b>	<b>123</b>
	<b>Appendix C Interactive Sample Size Calculation for Differential Gene Expression Experiments</b>	<b>125</b>
	<b>Appendix D Using Experimental Design for Data Processing and Visualization</b>	<b>127</b>



# List of Figures

2.1	Use of different gene expression technologies in recent years. . . . .	11
2.2	Using DNA microarrays to identify differential gene expression. . . . .	12
2.3	Using RNA-Seq to identify differential gene expression. . . . .	13
2.4	Comparison between a 2x2 and a 2x3 experimental design. . . . .	15
2.5	Mixture model t-distribution. . . . .	19
3.1	Software interfaces of the Experimental Design Wizard portlet. . . . .	31
3.2	Overview of the main functional requirements of the Experimental Design Wizard. . . . .	34
3.3	General features of the main use cases of the Experimental Design Wizard in BPMN. . . . .	35
3.4	Sample hierarchy graph. . . . .	40
3.5	Core components of the ISA data model. . . . .	41
3.6	qPortal Java libraries used by the Experimental Design Wizard. . . . .	44
3.7	UML diagram of the most important classes and libraries used in experimental-design-lib. . . . .	45
3.8	Flow diagram showing steps of the Experimental Design Wizard. . . . .	47
3.9	Sample source step of the Experimental Design Wizard. . . . .	47
3.10	Input of factor levels. . . . .	48
3.11	Tailoring step of the Experimental Design Wizard. . . . .	48
3.12	Input of factor levels for special factors. . . . .	49
3.13	Sample pooling function of the Experimental Design Wizard. . . . .	50
3.14	Registration step of the Experimental Design Wizard. . . . .	50
3.15	Metadata update table. . . . .	52
3.16	Adding experimental factors via metadata upload. . . . .	53
3.17	Collision display of metadata upload. . . . .	54
3.18	Import of ISA-Tab. . . . .	55
3.19	Completion of missing information after ISA-Tab import. . . . .	55
4.1	Graph aggregation principle. . . . .	63
4.2	Interoperability between qPortal, ISA-Tab and aggregation graph visualization. . . . .	64

4.3	Study design graph describing protein digestion and information about incomplete datasets. . . . .	66
4.4	Aggregation graph of a single ISA-Tab multi-omics study. . . . .	66
4.5	Comparison of study design graphs. . . . .	67
4.6	Visualization of an ISA-Tab study in the aggregation graph viewer after selecting “diabetes status” as the factor. . . . .	68
4.7	Visualization of the same ISA-Tab study after selecting the “age” as the factor in the viewer. . . . .	69
5.1	Runtime of sample size and power estimation functions provided by RnaSeqSampleSize. . . . .	76
5.2	Schematic view of the Sample Size portlet backend. . . . .	77
5.3	Presentation of datasets in the Sample Size portlet. . . . .	78
5.4	User interface for RNA-Seq sample size prediction portlet. . . . .	78
5.5	Comparison of FDR estimation using the OCPlus package and our portlet. . . .	79
6.1	Aggregation graph of the aspirin study. . . . .	88
6.2	Analysis of FDR for different sensitivity thresholds. . . . .	89
6.3	Selection of microarray QC workflow parameters. . . . .	89
6.4	Box plot and MA plots of normalized microarrays. . . . .	90
6.5	PCA plot and dendrogram of normalized microarrays. . . . .	90
6.6	Venn diagram of genes reported as differentially expressed using different methods. . .	91

## List of Tables

2.1	Example of a 3x2 full-factorial experimental design. . . . .	16
2.2	The null hypothesis and its connection to error types. . . . .	17
3.1	Definitions of concepts and terms used. . . . .	32
3.2	Adapted user interface terminology. . . . .	32
3.3	Mapping ISA to the openBIS data model. . . . .	42
3.4	Response times of the Experimental Design Wizard UI. . . . .	51
3.5	Studies created using the Experimental Design Wizard. . . . .	56
4.1	Aggregation graph creation time for different studies. . . . .	68

---

5.1	FDR estimations of our portlet based on OCplus. . . . .	80
5.2	Power estimations our portlet based on RnaSeqSampleSize. . . . .	80
6.1	Significantly expressed genes after multiple testing correction. . . . .	92
6.2	Functional annotation for differentially expressed genes between AS and AR groups. . . . .	93
B.1	ISA-Tab MetaboLights studies used for validation . . . . .	124
D.1	Significantly expressed transcripts found by Limma . . . . .	128



# Chapter 1

## Introduction

### 1.1 Motivation

Over the past two decades, the amount of data produced by the research of different parts of biomolecular systems has become increasingly massive. Repositories like the European Nucleotide Archive (ENA) administrated by the European Bioinformatics Institute (EBI) store petabytes of data, highlighting that biology has entered the big data age<sup>1–3</sup>. Most prominently, this data is related to efforts to analyze DNA sequences of different organisms in order to pinpoint genetic mutations causal to disease states or other phenotypes. In addition, the analysis of mRNA and protein differential abundance to learn more about the interplay of certain genes has been performed using microarray experiments, RNA-sequencing or through computational proteomics<sup>4,5</sup>.

Automation and more precise systems have empowered researchers to produce these data in an increasingly high-throughput fashion. Additional fields of study capitalizing on these developments are metabolomics, aimed at presenting an overview of the biological processing of different metabolites<sup>6</sup>, for example, given different environmental factors. In MHC ligandomics, computational proteomics methods are used to identify the small peptides presented by different cells of the immune system, allowing for new approaches for personalized cancer therapies<sup>7,8</sup>.

However, the dawn of the age of big data in life sciences has led to a number of open questions. Apart from storage issues, the large volume of data makes correct annotation critical if it is to be analyzed correctly and scientific standards of reproducibility are to be adhered to. Although the concept of metadata as data about data has been around since humans have begun organizing information, this construct has become increasingly important with the advent of the World Wide Web, and its collection and use has moved away from information professionals towards the general public<sup>9</sup>.

With respect to big biomedical data, annotation and analysis solutions need to scale well, as not only the volume of data is increasing, but the time it takes to generate it is steadily decreasing. Here, so-called context metadata—explaining what data, why and how it has been measured<sup>9</sup>—is crucial to facilitate automated data analysis steps. The variety of data make

comprehensive and adaptable data models necessary. Moreover, if different omics levels are to be integrated to utilize additional information, integration concepts are needed. In order to tackle these problems, there have been efforts to create standards for different types of experiments. MIAME, the Minimum Information About a Microarray Experiment<sup>10,11</sup> standard, and the microarray gene expression markup language MAGE-ML<sup>12</sup> aim at annotating experiments for sharing among researchers so they can be independently verified. Similarly, MIAPE, a standard describing the Minimum Information About a Proteomics Experiment tries to specify all the information necessary for interpreting proteomics experiments<sup>13</sup>. These standards are used to store experiment metadata in databases like PRIDE (PRoteomics IDentifications<sup>14</sup>). One of the latest approaches trying to incorporate these earlier solutions for different fields of study into an interoperable format for multi-omics is ISA-Tab, a spreadsheet format connecting meta information about research aims, different studies of an investigation and their assays<sup>15,16</sup>. Different efforts have been undertaken to provide users with software tools based on the ISA standard<sup>17</sup>, to leverage additional information from study factors and ontology information implicitly encoded in ISA-Tab and to connect it to different XML-based experiment formats<sup>18</sup>. Despite these efforts to standardize metadata collection, many recent articles have revealed large problems with both the reproducibility as well as the replicability of scientific studies<sup>19–22</sup>. Reproducibility can be defined as the ability to recompute results from existing data<sup>23</sup>. Issues to do so reliably can be caused by missing information about analysis parameters<sup>24</sup> or software tools used and success might be highly dependent on domain knowledge of the experimenter<sup>25</sup>. It can also be a problem of different software versions and the way they are installed or executed. Containerization of software pipelines and similar solutions try to tackle this problem<sup>26</sup>. However, using data from an existing study to come to the same conclusion is only part of the problem. To increase confidence in a scientific hypothesis, studies must be replicable, where replicability is defined as using the same experimental setup to generate new data<sup>23</sup>. Similar to the problem of missing annotation of analysis parameters, missing information about lab environments or the exact study protocol used can make this impossible. Principles like FAIR (Findable, Accessible, Interoperable and Reusable) have been developed to present guidelines on data and metadata collection and scientific data sharing<sup>27</sup>, but they have not yet been implemented widely by the community.

Other serious causes can be found in the experimental design approach itself. Blind studies and randomization are important tools to reduce biases. Power estimation and sample-size calculation are often overlooked, but crucial to put an experiment on a firm statistical footing<sup>28</sup>. Encouraging these standards is expected to help make research replicable<sup>29</sup>. This can also involve scrutiny of so-called “gold standards” on statistical approaches and suggestions for methods that might be more robust for high-throughput biomedical studies<sup>30</sup>. Guidelines that help tackle problems with both reproducibility, as well as replicability are needed to make data sharing feasible<sup>31,32</sup>.

In general, the new speed of biomedical data generation provides new opportunities for research, particularly with regard to experimental design. Since complex diseases are multi-causal and no perfect prediction of many disease states exists, genetic risk factors, as obtained from genetic loci in genome-wide association studies (GWAS<sup>33</sup>), can only predict a certain susceptibility to disease, pointing to other causes still to be determined. Owing to these circumstances, the paradigm of data-driven hypothesis-generation has suffered some setbacks. Results of many genetic variants initially reported to be related to disease states could not be reliably reproduced<sup>34</sup> and explanations for many heritable diseases are still incomplete<sup>35</sup>. Instances like this make it obvious that more data and robust experimental design is crucial to increase statistical significance. Apart from GWAS, full-factorial experiment designs can be used to further investigate interactions between multiple genes, environmental factors or both<sup>36</sup>, taking advantage of the large amount of available data that can now be generated relatively fast.

While the increasing wealth of big data in itself can help with better discriminating between signal and noise, it is important to ascertain the statistical power of a high-throughput experiment before performing it. Studies with too few individuals can lack statistical significance. At the same time, studies that investigate many more individuals than necessary can be costly in both time and money. Additionally, in the case of animal models, ethical concerns can arise. In the field of RNA sequencing, the choice between more replicates and deeper sequencing of samples has been widely discussed<sup>37–39</sup>, but is ultimately also dependent on the aim of a study.

To facilitate the planning phase of a study and be able to use data about the experiment for power estimation, data analysis or data sharing, researchers have to be persuaded to collect this information in a standardized form. In many cases, Excel spreadsheets are still the most widely used tool for research notes pertaining to assays and samples<sup>40</sup>. Most aforementioned standardization projects take this into account by providing spreadsheet-like metadata collection or experiment design formats. Despite this advantage in intuitiveness, spreadsheets are hard to curate and read by people unfamiliar with the study at hand. This is all the more problematic for large studies. Inconsistencies in the collected metadata are often not immediately obvious, especially since many standards allow missing entries or free text. In addition, conforming to these standards without guidance by software tools is hard. Installing the provided software is often hindered by different operating systems, versioning problems or simply convenience. One solution that has become popular in recent years is the use of science gateways providing these services platform agnostic through a web browser. Such portals try to overcome the disconnect between the different fields and parties involved in scientific projects and experiments. These research platforms are now an established approach and provide scientists with centralized interfaces to data, annotation, quality control and analysis tools that bring additional value to the project. To facilitate this, these portals must include intuitive and

well-documented integrated services that support researchers from experimental design to metadata collection and sample-size and power estimation. Here, we can define intuitiveness as enabling researchers to complete their tasks without extensive additional work in order to understand the software they are using. For the large studies of today's high-throughput medical science, this is all the more important, as the number of involved entities often prevents a naive approach to the representation or visualization of studies. This aspect also sets certain requirements for the speed of the involved tools.

### 1.2 Contributions of this Thesis

In this thesis, we present a web-based experiment design wizard, guiding researchers through the setup of biomedical experiments in order to create large omics studies with ease. Created study designs can be used to automate quality control and visualized using an aggregation graph based on study factors and the number of replicates. We further use registered study designs to analyze and visualize statistical power. For study planning, this approach is useful to estimate the needed sample size, dependent on the desired power. Pilot data can be used to estimate biological variation more accurately.

Chapter 2 covers the relevant technical, biological and statistical background. In Chapter 3-6 of this thesis, we describe contributions to the field of interactive study design, reproducibility, and automation of differential abundance analysis in computational biology.

1. We develop a model for experiment design and analysis, as well as an interactive interface for study design and integrate it into the science gateway qPortal. We implement import and update options for large and varied studies and provide interoperability with a widely-used metadata format, making our method a significant contribution to handle biomedical high-throughput experiments. We show the extensibility of our approaches using the example of proteomics and MHC ligandomics experiments.
2. We develop a visualization method for large study designs that aggregates samples based on their experimental factors. We explore the usefulness of this approach for our own, and for publically available studies.
3. We develop interactive interfaces to power analysis and sample size estimation methods for differential gene expression analysis that help researchers design their studies. The connection to our data model enriches these methods by providing users with stored pilot data and meta information about their experiment. This integration with our platform overcomes issues of other publically available web-tools and allows us to store analysis results in a way that conforms to FAIR data standards.

4. We apply the work of previous chapters by reproducing a microarray study on aspirin resistance. Building on our model for experimental designs, we create a quality control workflow that enables users to explore their raw data in relation to different experimental factors and view these results in the web browser. We compare the study's approach with knowledge gained from our interactive power analysis portlet.



# Chapter 2

## Background

This chapter outlines the biological, technological and statistical background of this thesis. Section 2.1 provides a brief overview of current technologies used to produce high-throughput biomedical data, putting a special focus on methods to measure differential gene expression. The related topic of experimental design and the statistical background is addressed in Section 2.2. In Section 2.3 we talk about the necessary characteristics of replicable science and about existing approaches to facilitate these needs.

### 2.1 Big Data and High-Throughput Omics Experiments

In recent years, technologies to measure and collect data about different aspects of biological systems have become increasingly faster, cheaper and more diverse. Perhaps the most prominent example of this process is the field of nucleotide sequencing, where the use of novel approaches made it possible to assemble a complete human genome at the turn of the millennium<sup>41</sup>. Shotgun sequencing, the sequencing of many random small fragments of DNA from a longer strand<sup>42</sup>, allowed to the Sanger chain-termination method to be used in a massively parallel manner. Paired-end sequencing, the practice of sequencing a fragment from both ends, provided more positional information of the fragments in order to help assemble the original genome with higher confidence<sup>43,44</sup>. While the cost of sequencing whole organisms, especially eukaryotes, was still prohibitive for wide application, the human genome project helped pinpoint the bottlenecks of the approach, soon leading to the development of Next-Generation Sequencing (NGS) technologies<sup>45,46</sup>. These high-throughput sequencing methods can be categorized by different library preparations, length of reads and concept of sequencing. The earliest and most common methods use a massively parallel approach of sequencing by synthesis. 454 pyrosequencing used emulsion PCR to amplify DNA into clonal clusters<sup>47</sup>. Each cluster was sequenced in its own sequencing well by triggering a reaction of the bioluminescent enzyme luciferase, once a nucleotide was added to the template sequence. Illumina sequencing follows a similar principle by sequencing amplified clonal DNA clusters on a flow cell<sup>48,49</sup>. Reversible terminator bases, that are labeled with different fluorescent dyes are added to determine the next complementary base. Afterward, unbound nucleotides and bound terminator bases are removed and the process repeats for the next position. Ion Torrent sequencing uses a similar

## 2. Background

---

synthesis approach but takes advantage of the release of hydrogen ions when nucleotides are incorporated into the DNA strand. A semiconductor is used to detect the resulting pH shifts<sup>50</sup>. Nanopore sequencing takes a fundamentally different approach by sequencing a single long molecule directly. By passing DNA through a narrow nanopore, different currents are induced by blocking ion flow<sup>51</sup>. The length of time of this blockage is dependent on the shape and size of the nucleotide, making sequencing possible.

Not only do these technologies lead to different read lengths, error rates and types, and the need for different analysis approaches, the number of protocols and applications is growing<sup>52</sup>. Apart from genome assembly using whole genome DNA sequencing, NGS today is used for variant analysis, transcriptomics (see Section 2.1.1), metagenomics and epigenomics through the study of DNA methylation (using Methyl-Seq or bisulfite sequencing)<sup>53</sup>. Chromatin immunoprecipitation sequencing (ChIP-Seq) can be used to explore DNA-protein interactions by identifying binding sites of DNA-associated proteins<sup>54</sup>.

The revolutionary advances of high-throughput methods in bioinformatics have not been limited to nucleotide sequencing and its applications. Besides NGS, one of the most notable technologies for biomedical experiments is the use of mass spectrometry for proteomics<sup>55</sup>. As with other analytical disciplines using spectrometry to disperse radiation or matter based on their properties, the aim of MS is to identify or quantify different spectrum components<sup>56</sup>.

MS outperforms the preceding methods, such as two-dimensional gel electrophoresis, in specificity and sensitivity, allowing not only the identification of different proteins, but even the standardised quantification of complete proteomes of simple organisms in a short time<sup>57</sup> and improved study of protein-protein interactions<sup>5</sup>.

Modern mass spectrometers ionize samples using one of a variety of methods<sup>58–60</sup>. For the field of proteomics, so-called “soft” ionization techniques are the methods of choice, as they impart less energy on the molecules in the sample, thus leading to fewer fragments. Matrix-assisted laser desorption/ionization (MALDI) accomplishes this by embedding the molecules of interest in a matrix, which is applied to a metal plate and absorbs energy from the laser<sup>61</sup>. For electrospray ionization (ESI), high voltage is applied to a liquid solvent to ionize the solute molecules. The ionized solution is accelerated towards an electrode. The repulsion between like-charged ions and the evaporation of the solvent causes a fine spray of gaseous ions<sup>62,63</sup>. This ionization step allows a mass analyzer in the spectrometer to separate different ionized molecules in the sample. Different mass analyzers have been developed<sup>64</sup>, with time-of-flight (TOF), quadrupoles and orbitraps being among the most prominently used techniques<sup>59</sup>. TOF MS uses an electric field of defined strength in order to accelerate molecules and measure the time until they reach a detector<sup>65</sup>. With molecules of larger mass being accelerated slower than lighter molecules, and highly charged ions accelerated faster than low charged molecules, the measurable mass-to-charge ratio ( $m/z$ ) provides a distinguishing quantity for characterization<sup>66</sup>. In a quadrupole mass analyzer, four metal rods are arranged

equidistant and parallel to each other. Ionized molecules are sent through the middle of the quadrupole. Through the application of voltage consisting of different parts of direct and alternating current, their trajectory can be changed so only ions with a specified  $m/z$  are able to pass the mass analyzer, while others collide with one of the quadrupole rods<sup>67</sup>. A detector then quantifies the number of ions leaving the mass analyzer, converting them into electrical signals. The abundances and unique mass-to-charge ratios of different ions lead to multiple peaks on the resulting mass spectrum, which is often represented in the form of a graph<sup>68</sup>.

For proteomics, a bottom-up approach is commonly used<sup>55</sup> to identify different proteins in a complex mixture. This strategy is also called shotgun proteomics<sup>69</sup>, referencing the shotgun sequencing approach discussed earlier. First, proteins are enzymatically digested into shorter peptides. In addition, liquid chromatography methods are often coupled with the spectrometer (LC-MS) to separate the resulting peptide mixture by hydrophobicity and charge<sup>70</sup>. After a first separation by  $m/z$ , ionized peptides are fragmented in a follow-up step, using an additional mass analyzer to separate and detect these fragments - a technique known as tandem mass spectrometry or MS/MS<sup>71</sup>. Here, mass analyzers such as the triple-quadrupole<sup>72</sup>, or hybrid methods like quadrupole time-of-flight (QTOF) can be applied<sup>73</sup>. The use of sequence databases then permits the search of resulting peptide mass spectra to match them to peptide sequences and subsequently infer the corresponding protein identifiers<sup>69,71</sup>. Besides preparatory steps including chromatography, methods like the previously mentioned gel electrophoresis also can be applied to reduce complexity and further increase resolution<sup>74,75</sup>. Other experimental approaches are available to determine the relative or absolute abundance of proteins or peptides: isotope and chemical labels like stable isotope labeling by/with amino acids in cell culture (SILAC)<sup>76,77</sup>, isobaric Tags for Relative and Absolute Quantitation (iTRAQ)<sup>78,79</sup>, Tandem Mass Tag (TMT)<sup>80</sup>, and others<sup>81,82</sup> are available to allow multiplexed measurements for quantitative proteomics.

As is the case for DNA and RNA modifications and the sequencing methods used to elucidate them, modern MS experiments also enable the study of posttranslational modifications (PTM) of proteins<sup>55,83,84</sup>. Studies of this type have focused on phosphorylation<sup>85,86</sup>, glycosylation<sup>87</sup>, methylation<sup>88</sup> and other modifications<sup>83,89,90</sup>. Some of these methods employ immunoaffinity purification, which uses antibodies recognizing the respective PTM in order to enrich proteins or peptides with that modification<sup>83</sup>. Notably, this strategy can also be used to enrich the processed peptides presented by different cells of the immune system, which vary by person, tissue, and in cases of various illnesses<sup>91,92</sup>. Named after the highly variable human leukocyte antigen (HLA), this field of HLA ligandomics enables new approaches in personalized medicine, cancer therapies or the creation of vaccines<sup>7,8</sup>.

It is easy to see that omics data share all the hallmarks of Big Data<sup>93</sup>. These characteristics are summarized using four (sometimes five<sup>i</sup>) V's. *Volume* refers to the amount of data that is

---

<sup>i</sup>the fifth V, business *value* is not a focus of this thesis

generated. *Velocity* refers to the speed at which data is generated and its mobility between different locations. As described, both the volume as well as the speed of data generation has massively increased. A single Illumina NovaSeq 6000 sequencing platform can produce data in the magnitude of around 1 terabase (Tb) per day<sup>94</sup>. It has been estimated that by 2025, between two and 40 exabytes of storage space will be needed for human genomes alone, eclipsing data generated in the field of astronomy, as well as on different social media platforms<sup>95</sup>. This leads to new and growing requirements for data transfer and analysis.

Another characteristic of big data, *variety*, refers to the different types of data that are generated. The variety of methods used to generate NGS or MS data and their numerous applications show the need for elaborate methods to handle the data<sup>52,53</sup>. Especially if we want to make assertions based on different approaches, as is often the case in research, information about the protocols and technologies our data was generated with must be readily available.

This is all the more important since different approaches bring with them different types of errors. For Illumina approaches, errors commonly lead to single nucleotide substitutions, while other methods can favor different nucleobases or lead to deletions<sup>96</sup>. Especially if we want to integrate information from different omics levels, different sources of errors can be problematic. Accordingly, the fourth V - *veracity* - refers to the quality or trustworthiness of the generated data. In this respect, quality control and repetition of our experiment are useful tools to make use of the wealth of big data in order to generate valuable knowledge.

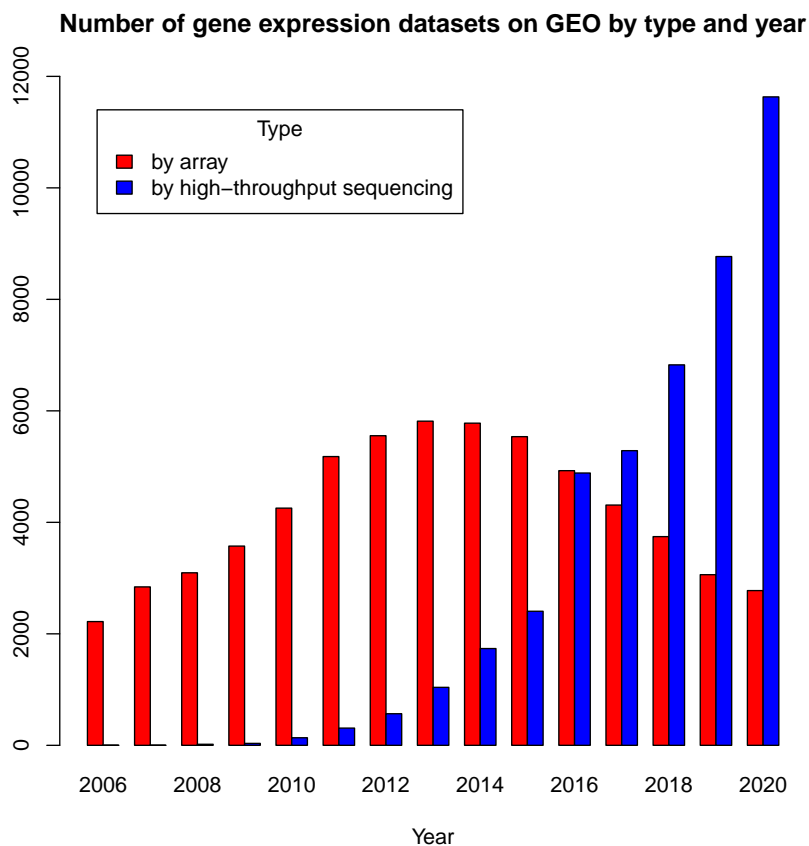
As we will discuss later, there are methods that can help us find the number of samples necessary to make significant statements about the outcome of our experiment. In order to do so, we first have to learn more about how our data is produced. Since our focus later will be on differential gene expression analysis, the focus of the following section will be on DNA microarrays and RNA sequencing (RNA-Seq).

### 2.1.1 Methods for Differential Gene Expression Analysis

Gene expression analysis has been an important topic of biomedical research for a long time<sup>97</sup>. While the genome itself can often help us understand variations between healthy phenotypes or mechanisms of disease, it is not always clear if these variants actually play a significant role in biological processes. Many genes are not expressed or only expressed in specific tissues<sup>98</sup>, which might not be related to the phenotype in question. Furthermore, the results of causative mutations can be better understood, if we can observe their effects on other parts of cellular processes<sup>99</sup>. Whereas mass spectrometry can elucidate expression changes in important protein families or the metabolism, DNA microarrays and RNA-Seq help explore such effects for messenger RNA (mRNA) before its translation into proteins.

Ease of sample preparation, relatively low cost and a general consensus on their analysis has made microarrays the platform of choice for many years, despite limitations like the difficult

detection of alternative splicing<sup>100</sup>. However, cheaper methods and a deeper understanding of RNA-Seq and its analysis have led to an explosion in the use of sequencing approaches. The number of publically available datasets in the Gene Expression Omnibus<sup>101</sup> database of each method (see Fig. 2.1) suggests that RNA-Seq has overtaken the use of microarrays for gene expression analysis in recent years. Nevertheless, both methods are still used to perform high-throughput experiments.



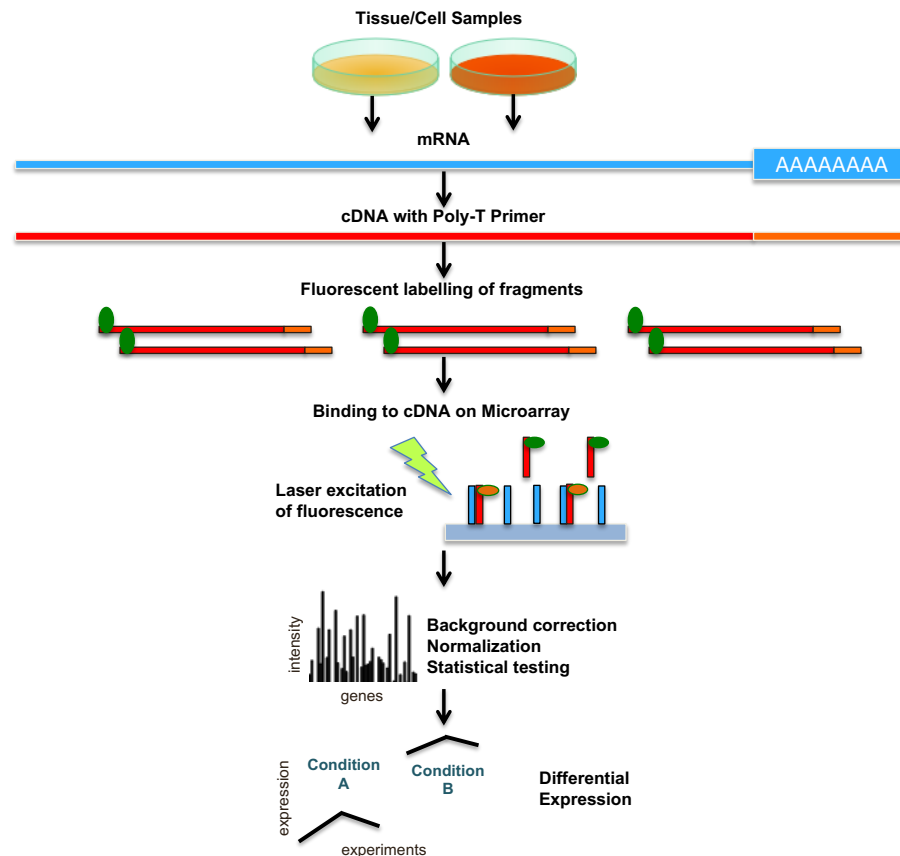
**Figure 2.1:** Number of datasets published each year to Gene Expression Omnibus (GEO). Search terms were the respective year and *expression profiling by array* or *expression profiling by high throughput sequencing* for the two respective dataset types. One dataset is defined as all available data for one study.

Both microarray and RNA-Seq experiments generally begin with the creation of a cDNA library<sup>102,103</sup>. For this purpose, mature mRNA containing a poly-A tail is extracted from cells. A poly-T primer is added along with free nucleotides and the enzyme reverse transcriptase (RT), which hybridizes the complementary DNA strand to the target mRNA. Afterward, the RNA strand is digested and DNA polymerase can be used to produce double-stranded cDNA,

## 2. Background

that can then be optionally amplified using methods like PCR (polymerase chain reaction)<sup>104</sup>. The resulting cDNA is then fragmented using enzymes.

In the case of microarrays (see Fig. 2.2), each fragment is labeled using a fluorescent dye<sup>105</sup>. cDNA fragments are then applied to the microarray chip. This platform consists of so-called probes, a set of oligonucleotides (in spotted microarrays) fixed to a solid surface that correspond to different, unique parts of genes.

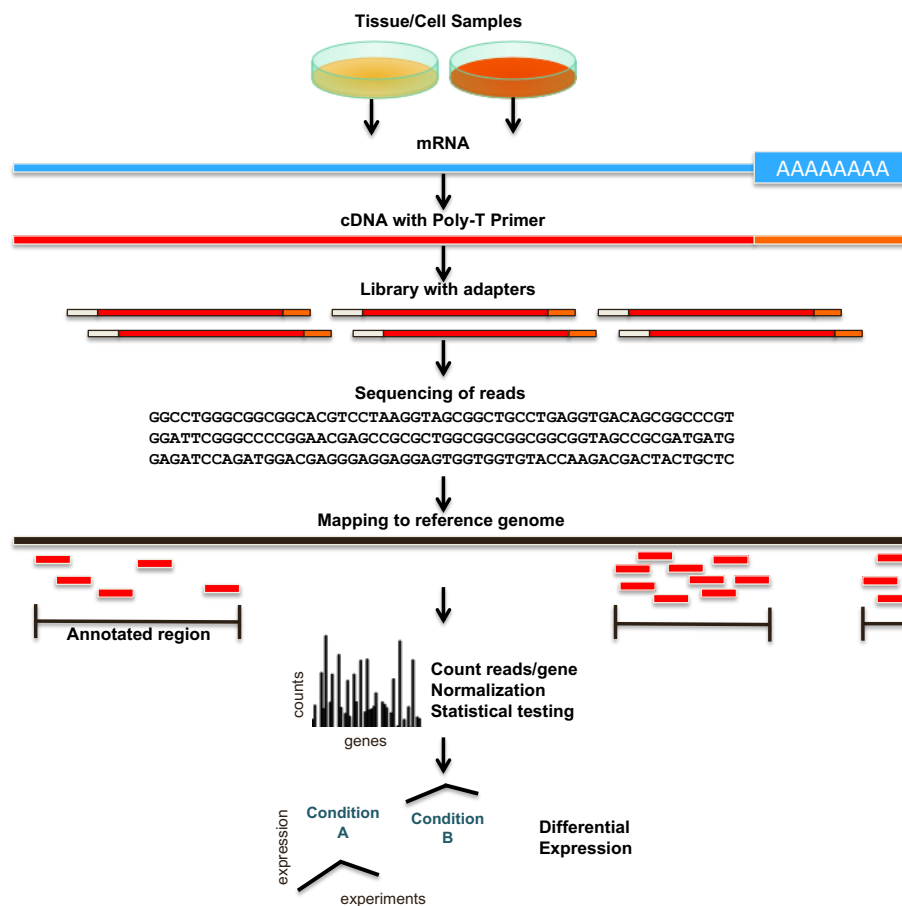


**Figure 2.2:** Schematic of the steps of a DNA microarray experiment. mRNA is extracted from cells under different experimental conditions and transcribed into cDNA. Fragmented cDNA is labeled with fluorescent markers. Known DNA probes on the microarray capture and hybridize complementary cDNA fragments of the experiment. Fluorescent light intensity is measured after excitation via laser and is stronger the more fragments are bound. After background correction and normalization, statistical testing is performed to identify significant differential expression between the testing conditions.

The target cDNA hybridizes with their complementary probes and all unspecific binders or non-binders are washed off. Excitation of the bound, labeled targets using a laser leads to a fluorescence effect. The light intensity corresponds to the amount of cDNA bound to a certain spot, thus enabling quantitative measurement of gene expression. The analog signal of light intensity is generally converted into a digital value. To compare different genes and

arrays, background correction and inter-array normalization are performed. Statistical testing can then pinpoint transcripts that are significantly differentially expressed between groups of experiments<sup>102,106</sup>.

For RNA-Seq using shotgun sequencing (see Fig. 2.3), index primers can be added to the cDNA fragments to allow for multiplexing, i.e. measuring multiple samples on the same sequencing machine<sup>107</sup>. Fragments are then sequenced using one of the methods mentioned in Section 2.1, typically resulting in millions of sequencing reads. These sequences are mapped to reference genomes using specialized alignment algorithms that can map the different transcript variants created by RNA splicing in eukaryotes<sup>108,109</sup>. Genome annotation makes it possible to count all reads mapped to a certain transcript or another region of interest.



**Figure 2.3:** Schematic of the steps of an RNA-Seq experiment. mRNA is extracted from cells under different experimental conditions and transcribed into cDNA. Adapters for sequencing and multiplexing can be added to cDNA fragments. The sequence of millions of fragments is determined in the sequencer and these reads are mapped to a reference genome sequence using software. Reads mapping to regions of interest on the genome are counted and can be normalized by gene size and the overall number of reads, depending on the application, before statistical testing can be performed to identify significant differential expression between the testing conditions.

## 2. Background

---

Different from the probes for DNA microarrays, which are standardized to the same length, RNA-Seq results are affected by transcript length and the overall number of reads. Widely used measures like the reads per kilobase per million mapped reads (RPKM, see Eq. 2.1<sup>110</sup>) — and the related fragments per kilobase per million mapped reads (FPKM) for paired-end sequencing — normalize the counted number of reads  $C$  mapped to a genomic feature by the total number of reads for a sample  $N$  and the length  $L$  in base pairs of the feature in question. This removes bias induced by transcript length and sequencing depth.

$$RPKM = \frac{10^9 \times C}{N \times L} \quad (2.1)$$

While RPKM and FPKM enable a relative comparison of expressed genes in one sample, normalizing by the number of reads of a sequencing run has the disadvantage of introducing a different bias between samples. Wagner et al.<sup>111</sup> show that since the number of transcripts can depend on the size distribution of RNA in the cell, samples of different tissues can lead to significant differences in RPKM despite similar expression. This is why measures incorporating read length and a better approximation of the number of samples transcripts have been proposed<sup>111</sup>. Many of the current tools for differential expression analysis of RNA-Seq data can, however, be used independently of these normalizations. DESeq<sup>2</sup><sup>112</sup> and edgeR<sup>113</sup> use underlying statistical models based on the raw read counts for each transcript.

## 2.2 Experimental Design

Based on Glass<sup>114</sup>, we can define scientific research as *the process of determining some property  $A$  about some thing  $R$ , to a degree of accuracy sufficient to be replicable by another person*. This is a broad definition that includes research like the description of a process, a species or an organ. As can be inferred from the beginning of this chapter, *descriptive science* plays a large role in high-throughput omics experiments, as it is often the foundation for many other experiment types. Without the determination of a reference genome or its annotation, there can be no experiment about causal variants or differential gene expression. Without an accurate description of the symptoms of a disease, it is hard to come up with a hypothesis about related cellular processes. Here, we will instead focus on the determination of cause and effect. This research type could be described as *the process of determining, if some property  $A$  leads to some outcome  $R$ , to a degree of accuracy sufficient to be replicable by another person*. This definition gives us two crucial aspects of experimental design: we need to observe the property we are interested in, as well as the outcome of the experiment. And we must make sure that our experiment can be verified by other researchers.

In the following, we will first describe how and why to incorporate multiple independent variables into one single experiment. Then we will discuss possibilities to predict how large our

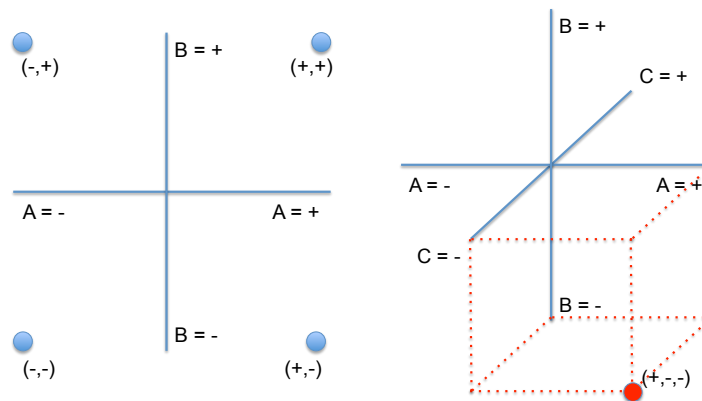
experiment should be, in order to satisfy the second part of our definition. As before, we will explain this process in more detail using the example of differential gene expression methods.

### 2.2.1 Factorial Design

Our definition of scientific research defines two variables. We call the outcome  $R$  of the experiment the dependent or response variable and  $A$  the independent variable, which is the focus of the study. In reality,  $A$  can be a highly complex mixture of traits and circumstances of our experiment. This means that a focus on a single variable that we want to investigate might lead to wildly varying results for us and anyone else trying to replicate our research. To counter this effect, multiple likely variables should be controlled for in a study. Ideally, this is done in a joint factorial experimental design instead of multiple designs.

The concept of factorial experimental designs, investigating multiple independent variables at once, was popularized in crop research<sup>36,115,116</sup>. Here, a *factor* is defined as one independent variable that is being studied. As previously described<sup>117</sup>, a *level* is one possible variation of a factor. The number of levels denotes the total number of different variations for a single factor that was used in an experiment.

Factorial designs are called full-factorial designs (fully cross-factored designs) if every possible combination of levels is tested<sup>118</sup>. Otherwise, they are incomplete, or unbalanced, factorial designs. Representations of complete factorial designs are shown in Figure 2.4 and Table 2.1.



**Figure 2.4:** Visual representations of full-factorial experimental designs with two levels (- and +) for each independent variable. Left: experimental design testing two independent variables  $A$  and  $B$ , leading to four unique testing conditions, shown as circles. Right: experimental design testing three independent variables  $A$ ,  $B$  and  $C$ . The eight unique testing conditions, of which one is shown as a red circle, can be located at the vertices of a cube.

A quantitative variable can be measured as an ordinal or interval (e.g., age), while qualitative or categorical variables are measured on a nominal scale, for example as a disease state.

## 2. Background

---

**Table 2.1:** Example of a 3x2 full-factorial experimental design. Two variables  $a_1$  and  $a_2$  containing three levels each (−, +, 0) are tested, leading to nine different experiments.

Variables	Experiment no.								
	1	2	3	4	5	6	7	8	9
$a_1$	−	−	−	+	+	+	0	0	0
$a_2$	−	+	0	−	+	0	−	+	0

One benefit of performing factorial design experiments is that they save resources<sup>118,119</sup>. More importantly, testing multiple factors at once allows experimenters to detect interactions, which is not possible in one-factor-at-a-time (OFAAT) experiments.

### 2.2.2 Statistical Power

While the type of experimental design is crucial in order to obtain a significant result, considerations about sample sizes and statistical power have to be taken into account before measuring data, as well. There is often no statistical fix for errors in planning, once an experiment has been performed.

In essence, any study comparing the outcome of experiments based on a factorial design tries to solve the problem of finding a significant correlation between one or more factors  $A_n$  and a response variable  $R$ , data of which is collected by observing different levels of the factor. This is complicated by known or unknown confounding factors  $C$ , which also have an effect on the response variable<sup>118</sup>:

$$R = A_1 + A_2 + C \quad (2.2)$$

The (statistical) power of an experiment to show that one or more factors are indeed correlated with the response (i.e. the probability to reject the null hypothesis  $H_0$  if the alternative hypothesis  $H_1$  is true (see Table 2.2)), is itself dependent on multiple factors: the effect size of  $A$  on  $R$ , the amount of random variation introduced by confounding factors and the number of replicates observed for each level of  $A$ .

There are multiple strategies that make use of this knowledge<sup>118</sup>. In order to limit noise, researchers can try to control for known confounding factors, which is one reason for the use of multi-factorial designs. Other approaches include blocking or matching. Here subjects are split into groups while keeping confounding factors like the age of subjects similar between the groups. To control for unknown confounding factors introduced by the observation process of subjects or samples itself, experiments are often measured in a randomized order and treated in the same manner, apart from the variation of study factor levels.

Increasing the number of replicates can have a large effect on statistical power since life scien-

**Table 2.2:** The null hypothesis is the default hypothesis stating that no effect has been observed in an experiment, here in respect to the study of factor A and response R. Dependent on the true effect of A on R and the outcome of a study, two different types of errors can be made. The probability of type II errors decreases with the statistical power of a study design.

Reality	Study Result	
	Relationship between A and R	No relationship between A and R
$H_0$ : There is no relationship between A and R	Type I Error	Correct
$H_1$ : There is a relationship between A and R	Correct	Type II Error

tists study highly complex systems with many confounding factors they have little control over. Here, pseudo-replication, like the observation of closely-related subjects, should be avoided, as genetic similarity introduces bias, which can lead to type I errors<sup>120</sup>.

While the effect size of a factor cannot be manipulated in a planned fashion, researchers can use it to specify a lower bound of the effect that they wish to predict<sup>118</sup>. Reasons for this can be cost or health considerations. For example, a new drug being developed in drug discovery needs a to show a minimum effect size in order to be approved, especially if there are existing treatments available. Thus it would not make sense to increase the number of replicates to detect smaller effect sizes.

Knowledge or estimations of the respective parameters can be used to compute either the statistical power or the needed number of replicates before an experiment is performed. Both effect size and biological variation can be found or estimated if the system being studied is well-known<sup>118</sup>. For example, natural variation in the lifespan of mice can be easily computed and the magnitude of a drug's effect on longevity might be known. In the case of more complex types of experiments, more elaborate mathematical models are often used to predict variation based on the data of a pilot experiment or on domain knowledge collected from similar experiments.

Differential expression analysis is one typical example of modern high-throughput experiments used to discover genes related to disease or other phenotypes. Raw data of the abundance of genetic transcripts can be measured by both DNA microarrays as well as RNA sequencing (RNA-Seq). These methods follow different protocols and statistical assumptions and thus different approaches are needed in order to compute statistical power. These differences are described in more detail in the following sections.

### 2.2.3 Statistical Power of Microarray experiments

Microarray analysis to measure differential expression of genes typically follows a hypothesis-driven statistical analysis<sup>121,122</sup>. Gene expression changes are declared to be statistically significant between two groups if a test-statistic shows that their means are not equal<sup>123</sup>. Since microarray data is continuously distributed, it is usually fit to the normal distribution or related probability distributions, often using the standard or student's t-test to detect significant differential expression<sup>121,122</sup>. This procedure is based on selecting a critical value, which corresponds to the type I error rate  $\alpha$  of declaring one non-DE gene as differentially expressed.  $\alpha$  is also known as the significance level<sup>123</sup>. However, since the common use case for DNA microarrays is to measure thousands of genes, this approach leads to the multiple testing problem<sup>124</sup>. While  $\alpha$  may be adequate for single statistical tests, the expected number of type I errors, as described by the false discovery rate (FDR), can become prohibitive when studying whole genomes. The family-wise error rate (FWER)  $\bar{\alpha}$  (see Equation 2.3) indicates the probability of rejection of at least one true null hypothesis. For the typical threshold of  $\alpha = 0.05$  and only  $m = 50$  tested genes, the chance for one such false positive is already 0.92.

$$\bar{\alpha} = 1 - (1 - \alpha)^m \quad (2.3)$$

There are multiple procedures to correct for multiple testing. The Bonferroni correction<sup>125,126</sup> tightens the significance level by dividing  $\alpha$  by the number of performed tests, guaranteeing an  $\text{FWER} \leq \alpha$ . A less strict modification is the Holm-Bonferroni method<sup>127</sup>, for which p-values  $P_{1 < k < m}$  of  $m$  hypothesis tests are ranked and the tests with lowest p-values are rejected until

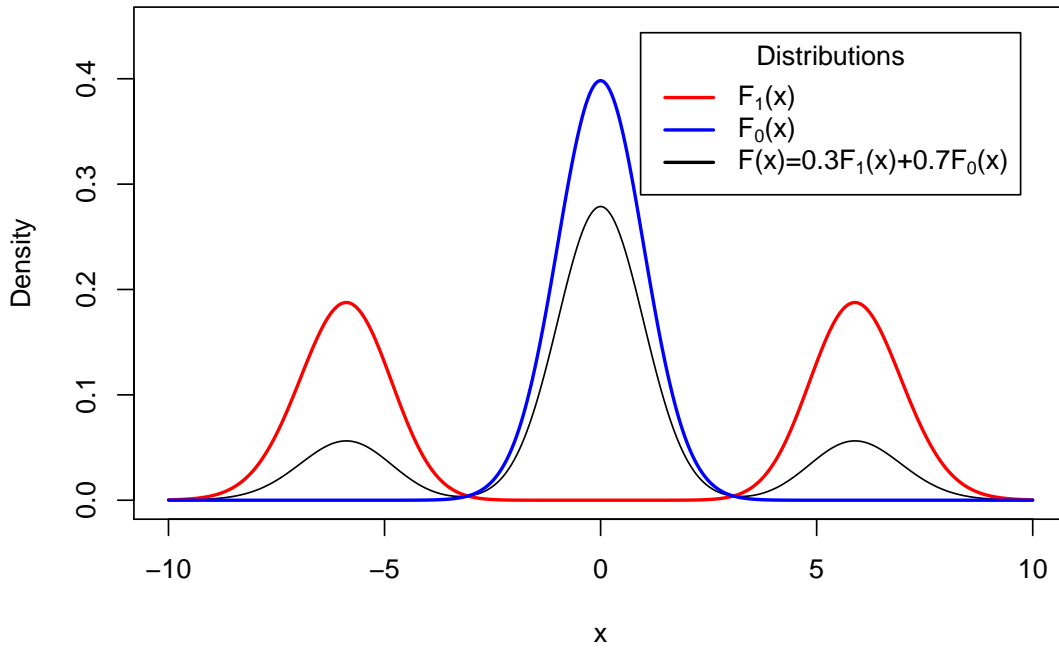
$$P_k > \frac{\alpha}{m + 1 - k} \quad (2.4)$$

While all of these methods try to minimize type I errors, they can be problematic with respect to type II errors. This means they might not reject the null hypothesis for genes that are in fact differentially expressed, increasing the false negative rate (FNR). For this reason, Pawitan et al.<sup>128</sup> argue that the question of statistical power in Microarray analysis should rather be considered a classification problem. Controlling for FDR and FNR has the advantage of taking into account factors such as the ratio of expected DE genes to all genes as well as the effect size that researchers want to be able to detect in their experiments. In fact, using the correct model, we can control for both types of errors at the same time when declaring a percentage of genes as DE. In the following, we illustrate in more detail how this has been done using the mixture model approach of Pawitan et al.<sup>128</sup>.

Figure 2.5 shows an example of the mixture Student's t-distribution defined by

$$F(t) = p_0 F_0(t) + p_1 F_1(t) \quad (2.5)$$

where  $p_0 F_0(t)$  is the central t-distribution scaled by the proportion of truly non-DE genes  $p_0$  in the dataset and  $p_1 F_1(t)$  is the sum of non-central t-distributions signifying positive and negative fold change of truly DE genes in the dataset, whose proportion is  $p_1 = 1 - p_0$ .



**Figure 2.5:** Mixture model t-distribution consisting of a central t-distribution scaled by 0.7 and two non-central t-distributions scaled by 0.15, each, which signify a log-fold change of -1 and 1, respectively.

For a critical value  $c > 0$  we can compute a number of statistical measures for tests on the whole dataset:

- Proportion of declared DE genes:

$$2(1 - F(c)) = 2F(-c) \quad (2.6)$$

- False discovery rate:

$$FDR = \frac{p_0(1 - F_0(c))}{1 - F(c)} = \frac{p_0 F_0(-c)}{F(-c)} \quad (2.7)$$

## 2. Background

---

- Sensitivity, power or true positive rate (TPR):

$$TPR = 2(1 - F_1(c)) = 2F_1(-c) \quad (2.8)$$

From Equations 2.7 and 2.5 follows the relationship between FDR and sensitivity:

$$\begin{aligned} 1 - FDR &= \frac{F(-c) - p_0 F_0(-c)}{F(-c)} \\ (1 - FDR) \times F(-c) &= p_1 F_1(-c) \\ \frac{(1 - FDR) \times 2F(-c)}{p_1} &= 2F_1(-c) = TPR \end{aligned}$$

If we declare a ratio of  $p_1 = 2F(-c)$  genes as DE, we can control for both FNR as well as FDR of our test at the same time:

$$TPR = 1 - FDR \Rightarrow FDR = 1 - TPR = FNR \quad (2.9)$$

This means that the resulting statistical measures depend only on the mixture of distributions we are observing, which are defined by their degrees of freedom  $df$  and the non-central distributions. The degrees of freedom of the t-distribution are based on our sample size. The non-centrality parameter is based on sample size and the log fold change we want to detect.

### 2.2.4 Statistical Power of RNA-Seq experiments

As a count-based measure, RNA-Seq expression cannot be modelled using the continuous normal distribution. Instead, analysis tools like DESeq<sup>112</sup> and edgeR<sup>113</sup> use the discrete negative binomial distribution (also known as gamma-Poisson distribution). Here, within-group variability of the experiment is described as the variance of read counts for gene  $i$  and replicate  $j$ , which is dependent on the mean of read counts  $\mu_{ij}$  and the dispersion  $\alpha_i$ <sup>112</sup>:

$$Var_{ij} = \mu_{ij} + \alpha_i \mu_{ij}^2 \quad (2.10)$$

Since variance is an absolute measure, it is unsuitable for comparisons between genes with different magnitudes of read counts. The coefficient of variation  $CV = \frac{\sigma}{\mu}$  measures the relative standard deviation and can be derived as follows:

$$\sigma^2 = \mu + \alpha\mu^2$$

$$\frac{\sigma^2}{\mu^2} = CV^2 = \frac{1}{\mu} + \alpha$$

The CV is dependent on both the mean and dispersion. Since the mean read counts increase with sequencing depth of a study's samples, dispersion  $\alpha$  reflects the true biological variation between replicates as well as any errors in library preparation<sup>129</sup>.

This makes the estimation of  $\alpha$  a crucial step in reporting differential expression as well as determining the needed sample size of an RNA-Seq experiment before it is performed. Dependent on the method and aim of a study, the dispersion can be approximated for each gene or as a single parameter describing in-group variability of all genes in a study. It is common for most methods to pool information across many genes in order to estimate dispersion<sup>112,113,130,131</sup>. For DESeq2, Love et al.<sup>112</sup> use a maximum likelihood estimation to produce raw dispersion values for each individual gene. A smooth curve is fitted to describe this population of different dispersion values with respect to the mean of normalized counts. Finally, a shrinkage estimator is used to compute the maximum *a posteriori* (MAP) estimation for each gene  $i$ , moving raw dispersion values closer to the curve:

$$\alpha_i^{MAP} = \operatorname{argmax}_{\alpha}(\alpha_i^{gw} + \Lambda_i(\alpha)),$$

with the penalty term:

$$\Lambda_i(\alpha) = \frac{-(\log(\alpha) - \log(\alpha_{tr}(\bar{\mu}_i)))^2}{2\sigma_d^2}$$

Here, the gene-wise dispersion estimate  $\alpha_i^{gw}$  is obtained by maximizing the Cox-Reid-adjusted likelihood<sup>132</sup> of the dispersion. The penalty term compares prospective dispersion values  $\alpha$  with the trend line (fitted curve) of dispersion estimates over the means of the normalized counts  $\alpha_{tr}(\bar{\mu}_i)$ .

$\sigma_d^2$  is the prior variance, which is dependent on the degrees of freedom of the sampling distribution, i.e. the number of replicates and conditions used in the experiment. Resulting, dispersion values are moved further in direction of the fitted curve, the less evidence, in the form of the amount of samples, is provided<sup>112</sup>. Additionally, outliers with low predicted gene-wise dispersion are moved closer to the curve, while outliers with very large predicted dispersion are not altered in order to prevent false positives.

### 2.3 Reproducibility and Replicability

Reproducibility and replicability have been discussed in the past decades in the context of a number of different scientific fields such as geophysics<sup>133</sup>, epidemiology<sup>134</sup> and psychology<sup>135</sup>; sometimes using varying definitions<sup>23</sup>.

In the following, we will use the older and more common notion<sup>23</sup> of defining reproducibility as the ability to recompute existing data analysis results<sup>22</sup>.

To guarantee reproducibility, certain information must be available to investigators:

- Raw data from the experiment
- Algorithms that have been run on the data and their respective versions
- Used parameters, including a correct mapping for input files
- The same execution environment

The basis of reproducing experimental results of an earlier study is the complete data that has been measured. Raw data is defined as the data that has been derived directly from the measurement apparatus, be it a DNA sequencer or manual measurements written down by a scientist<sup>32</sup>. Data that has been modified in any way is not raw data and can slow down processing or even make reproducing a study impossible if intermediary steps are not provided. In this respect, accompanying metadata is also crucial to reuse raw data, as it is necessary to know how data has been generated. In addition, files have to be uniquely relatable to their study variables, and often even to individual replicates.

Similarly, the used analysis algorithms, their versions, and their input parameters are important. Results of data preprocessing steps, like normalization or genome mapping are dependent on the used software and parameters. Genomic variant calling pipelines lead to vastly different results dependent on the type of sequence data used and on the variant calling tool<sup>136</sup>.

The execution environment is important, as running data analysis on different computational platforms can lead to numerical instability<sup>137</sup>. While the resulting differences are typically small, they can be a major problem for result verification.

Reproducibility, however, is only one part of good research. Since raw data is inevitably biased towards the circumstances of its creation, completely independent verification of experiments is required. Ultimately, science needs to be replicable to find out if any predicted effect size is larger than any introduced bias. Here, replicability is defined as the ability to produce a consistent result when performing an independent experiment targeting the same scientific question.

To guarantee full replicability all reproducibility criteria must be fulfilled. Additionally, the

experiment must be statistically sound, as insignificant results are by definition unlikely to be replicable.

Current approaches that try to solve these problems to facilitate efficient data sharing, reproducibility and replicability are described below.

### 2.3.1 Standards for Experiment Reporting

The key to successfully communicate the necessary experimental design information and additional study metadata is using a *language* that everyone involved in this research understands. Multi-omics experiments have complicated this task in the regard that experimental designs have become more diverse and the analytical techniques used are more complex<sup>15</sup>.

Early initiatives for standardized experiment reporting were either driven by regulatory frameworks and agencies like the FDA (Food and Drug Administration)<sup>138,139</sup>, by scientific journals<sup>140</sup> or by consortia focusing on particular technologies<sup>141</sup>. In the latter case, the focus was on providing tool interoperability and data exchange by demanding common minimal requirements. Some of these established checklists are the MIAME<sup>10,11</sup> standard for microarray experiments, the MIAPE standard for proteomics experiments<sup>13</sup>, the MIMIx standard for molecular interaction experiments<sup>142</sup> and the MIRAGE standard for glycomics experiments<sup>143</sup>. Many journals require that authors comply with these standards when sharing their experimental data, but the isolated development of these minimum information checklists has led to various problems and made it difficult to establish the full range of minimum information standards. The Minimum Information for Biological and Biomedical Investigations (MIBBI) project aims to facilitate coordination in order to develop an integrated checklist resource for the wider bioscience community<sup>144</sup>.

While information checklists define the minimum requirements for reporting experiment metadata, other initiatives have been focused on developing common data models and syntax to make studies interoperable on all levels. Many approaches are again driven by different omics technologies as well as the intended use of the format. XML-based models enable easy specification and verification of complex metadata objects. MAGE-ML<sup>12</sup>, which is based on the MIAME standard for microarrays, and mzML<sup>145</sup>, which supports the MIAPE standard, enables interoperability with different software tools, while still maintaining basic human readability of the described information. Their tabular format counterparts like MAGE-TAB<sup>40</sup> and mzTab<sup>146</sup> often target researchers who are most familiar working with Microsoft Excel or who want to perform downstream analysis on sets of genes, transcripts or proteins using software like R. While these formats can support minimum information standards, they are often additions to the more complete XML-based formats and primarily serve as an easily accessible summary of the most important data and metadata.

## 2. Background

---

A common problem of these technology-based data standards becoming obvious once multiple omics levels are measured for the same study is redundancy. Since biological source material, sample processing, and experimental factors are common and important parts of the experimental design, they would need to be reported for each of the different standards of each of the technologies used to study a sample. This is why efforts like the Functional Genomics Experiment data model (FuGE) have been undertaken to facilitate convergence of different standards for high-throughput experiments in biology<sup>147</sup>.

While FuGE goes the way of providing a complex data model, the ISA-Tab standard tries to generalize the user-friendly MAGE-TAB format and use its syntax for a wide variety of biomedical experiments<sup>15,16</sup>. ISA-Tab splits information about a project into the sub-parts investigation, study, and assay. Studies, so-called units of research, contain information about subjects, their sources, characteristics, and treatments. Subjects of each study can be analyzed using different assays to perform analytical measurements that lead to different types of data, for example gene expression. The assay part also connects each file of a project to the relevant meta information. The investigation file defines the project context: declarative information used in studies and assays and connects these files. Despite being based on the MAGE-TAB standard, ISA-Tab does neither enforce the MIAME checklist for microarray experiments nor any other minimum information requirements, letting users implementing the standard decide on how to regulate its use. ISA-Tab has become the basis for several further developments: the ISA software suite assists users with metadata annotation, enabling the use of checklists and ontologies, and facilitates submission of experiments to public repositories<sup>17</sup>. linkedISA provides a translation into a more complex data model in the form of a semantic web representation of the ISA-Tab syntax. This also allows extraction of additional implicit information like experimental factors from the format<sup>18</sup>.

The growing number and sophistication of different reporting standards has raised questions about the most important characteristics needed to enable scientific reproducibility and replicability. The FAIR (Findable, Accessible, Interoperable and Reusable) principles for scientific data management and stewardship have the intent to act as guidelines on data and metadata collection and sharing, in order to facilitate scientific data sharing<sup>27</sup>.

The guiding principles state that data and metadata must be *findable*<sup>27</sup>. A prerequisite for this is the use of unique and persistent identifiers. Of course, searchability must also be applicable to any attributes of the data in question. It is therefore important that data are described with rich metadata. To facilitate efficient searching, data should be indexed or registered.

Data must also be *accessible*<sup>27</sup>. This means standardized communications protocols need to be used to not make data access needlessly complicated. Open and free protocols that can be implemented as part of different tools are needed. Depending on the type of data,

authentication and authorization procedures must be available. While data does not have to be stored indefinitely, metadata must stay available.

The third guiding principle, *interoperability*<sup>27</sup>, specifies that metadata must use a formal, accessible, shared, and broadly applicable language. To this end, the vocabularies to annotated data should also follow FAIR principles and metadata should reference other metadata.

The FAIR principles also define rules for data *reusability*<sup>27</sup>, such as the inclusion of a well-defined license for data usage. The data provenance, which specifies in detail which steps were used to transform the raw data, must be available and associated with data and metadata. Last but not least, data and metadata must meet the domain-relevant community standards, in order to enable true reusability.

Standardizing experiment reporting the correct way does not only enable reproducibility and replicability, but it can also allow for the automation of different steps of the study, speeding up the overall analysis and eliminating human error. In order to evaluate the FAIRness of processes and systems, there have been efforts to define metrics for compliance with the FAIR principles<sup>148–150</sup>.

### 2.3.2 Science Gateways and Workflows

Some content of this subsection is part of the manuscript:

---

#### *qPortal: A platform for data-driven biomedical research*

Christopher Mohr<sup>+</sup>, Andreas Friedrich<sup>+</sup>, David Wojnar, Erhan Kenar, Aydin Can Polatkan, Marius Cosmin Codrea, Stefan Czernel, Oliver Kohlbacher, Sven Nahnsen *PloS ONE* 13.1 (2018)

+ These authors contributed equally

---

Whereas standardized experiment reporting can provide important guidelines for data sharing, the means of metadata collection for experiments and analysis in itself cannot be overlooked. The complexity of modern biomedical experiments often necessitates the involvement of multiple collaboration partners from different fields, working at different locations. All of them must have access to their project data in order to annotate it.

In addition, centralized solutions, which are becoming more common as research consortia collect data on a large scale, introduce hurdles on the infrastructure and computational side. Examples are ensuring data access, availability of analysis pipelines for specific omics data, and data security, especially with respect to clinical data. Web-based science gateways can bridge the gap between the different parties involved in scientific projects. Such platforms are an

## 2. Background

---

established approach to provide scientists with a centralized interface to data, metadata and analysis tools that bring additional value to the project.

In recent years, efforts that try to tackle these problems by developing portal-based solutions have been undertaken for specific biomedical research fields. These range from solutions for proteomics<sup>151</sup>, genomics<sup>152</sup> to portals providing analytical tools for the analysis of multiple omics levels<sup>153</sup>. Other web-based solutions focus on specific areas like cancer research<sup>154</sup>, neuroscience<sup>155</sup> or phylogenetic analyses<sup>156</sup> and can even include a complete Laboratory Information Management System (LIMS)<sup>157,158</sup>.

Prerequisites for successful science gateways are standardized web-frameworks like VAADIN that support a wide range of browsers<sup>159</sup>. Furthermore, many science gateways are based on different web portals like Liferay or JBoss that offer basic functionality to handle users and security<sup>160</sup>. On top of these portals, different tools can be run as portlets. Portals that implement the Java Portlet Specification (JSR 168) enable compatibility of these portlets across different platforms<sup>161</sup>, making collaboration easier.

Typically, these portal solutions contain interfaces to workflows systems, enabling users to analyze or transform data in predefined ways, each step symbolized by one workflow node. Since workflows handle input, output, and communication between different nodes, they can be useful for researchers not familiar with command line tools or even certain steps in a processing pipeline, given that an intuitive interface to the respective workflow is provided. Furthermore, distributed resource managers (also known as job schedulers) like Moab<sup>162</sup>, TORQUE<sup>163</sup> or Slurm<sup>164</sup> queue workflow instances and distribute the workload to different compute resources.

Science gateways and workflow systems are important tools to facilitate reproducibility, as they often provide easier ways to investigate data provenance, than when individual tasks of data transformation are performed by researchers. However, they cannot provide a perfectly reproducible environment on their own, especially if they are set up at different facilities where different versions of software dependencies or even of the analysis tools themselves might be installed.

### 2.3.3 Virtual Machines and Containers

To truly standardize data analysis and enable data sharing between different facilities, additional steps have to be taken. Virtualization is a concept that was first adopted on a large scale when servers became powerful enough to run multiple applications, making it possible to consolidate many specific servers into a single system<sup>165</sup>. Here, a hypervisor, an interface that is running directly on the server hardware, divides the physical hardware for multiple virtual machines (VM). Each VM consists of its own operating systems (OS) as well as specific applications. The hypervisor translates each I/O request from a VM operating system to the real-world storage and provides the response to the virtual OS. Memory or CPU processing is handled accordingly.

This enables more efficient use of resources, while at the same time providing security for the host system and the different virtual machines since the latter are not aware of each other and the hypervisor controls all interactions with real physical hardware.

Containers are a similar concept that represents a lightweight, more efficient version of virtualization<sup>166</sup>. Unlike virtual machines, containers share a common OS kernel and other components. This makes size, startup time and maintenance of containers more easily manageable. For these reasons, container solutions like Docker<sup>167</sup> and Singularity<sup>168</sup> have been proposed for fast prototyping and sharing of high-throughput analysis pipelines.

While the shared components of the operating system cannot be changed by the container, several container solutions have security drawbacks, since direct interaction with the host system is not controlled via hypervisor<sup>169</sup>.

Where containers can provide speed and convenient sizes for sharing, virtual machines can provide a second layer of safety and standardized runtime environment. This makes both technologies a good match in order to further increase security and reproducibility of workflow systems as they are used as part of biomedical research platforms<sup>137</sup>.



## Chapter 3

# Modeling of Experimental Designs

Some content of this chapter is part of the manuscript:

---

*qPortal: A platform for data-driven biomedical research*

Christopher Mohr<sup>+</sup>, Andreas Friedrich<sup>+</sup>, David Wojnar, Erhan Kenar, Aydin Can Polatkan, Marius Cosmin Codrea,

Stefan Czemmél, Oliver Kohlbacher, Sven Nahnsen *PloS ONE* 13.1 (2018)

+ These authors contributed equally

---

### 3.1 Introduction

Subpar experimental design and missing metadata annotation are seen as prime reasons for the so-called reproducibility crisis in science<sup>170</sup>. Community efforts have been taken to simplify metadata acquisition and re-use by providing standardized formats<sup>11,13,16</sup>. In order to involve a broader scientific audience, these standards often build on storing experiment and sample information in spreadsheet formats, as they are commonly used by experimenters in the lab. While generally being easy to understand and work with, the size and complexity of modern biomedical experiments call for methods that help researchers collect and check the necessary information. Different efforts have been undertaken to provide users with software based on the ISA standard<sup>17,18</sup> in order to enable ontology support and sanity checks. These tools mostly take the approach of displaying sample and assay information in tables, which provides a complete overview of all available information, but does seldom simplify the common problems of large studies. Web-based science gateways enable automatic metadata collection or provide users with intuitive, platform-agnostic means of creating and analyzing experiments<sup>151–153</sup>.

Nonetheless, some of the most widespread solutions often focus on workflow annotation — a necessary, but not sufficient step of performing reproducible experiments. While many web platforms offer a wide variety of additional annotation features, it is often assumed that researchers have already planned a well-defined experiment, or — in the worst case — even collected data. Another assumption is that they will know what meta information to

add to guarantee a reproducible or replicable experiment. Creation of projects online or of shareable study formats is consequently seen as mere metadata collection. This uncoupling of experimental design and experiment creation presents a missed opportunity to guide researchers to statistically sound study design. Starting with extensive metadata collection before the experiment has numerous advantages for scientific studies: mistakes in study design or sample handling can be traced back more easily and with higher confidence. Time and money can also be saved because the study design allows for the estimation of statistical power before experiments are performed. In addition, well annotated experimental data is much more likely to be reused in future research. Studies supporting this approach note that without sound experimental design even computationally reproducible results have to be used with caution<sup>22</sup>. To increase confidence in a scientific hypothesis, studies must also be replicable when using the same experimental setup to generate new data.

While guiding users through the process of experimental design and collecting metadata before the experiment provides clear benefits, it can pose new problems that must not be overlooked. Not all information about a study can be known before its inception. Methods that allow metadata annotation must present easy means of adding information that is collected during or after the measurement of samples, or they risk losing annotation due to missing usability. Especially for projects with many replicates and independent variables, factorial designs can lead to a large number of cases that can make it hard for researchers to keep the stored information consistent with their experiments. Methods are needed to conveniently manage and update these large studies.

Here, we present our work on a data model to describe experimental designs and related meta information. We developed a web-based wizard to guide researchers through the process of creating a full factorial experimental design. We implemented different entry-points to enable a heterogeneous user group to interact with existing designs to add more metadata or samples. Our solution provides options to import large studies and interoperability with the ISA-Tab format. We show the extensibility of our model and software tool for different technologies and investigate its compliance with principles of reproducibility.

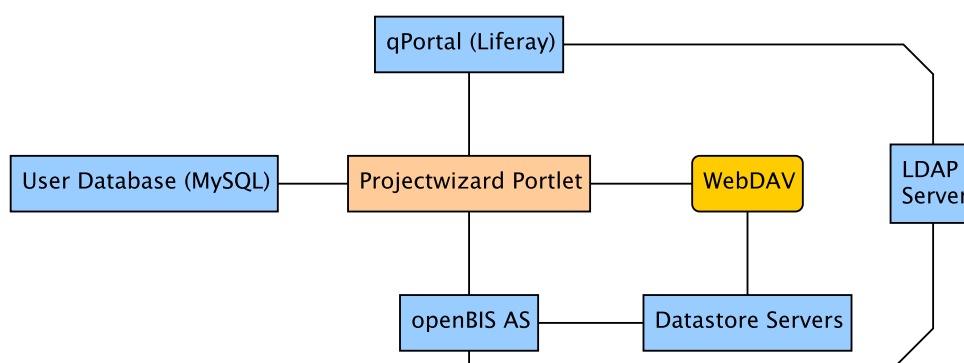
## 3.2 Requirements

The purpose of our web application is to enable researchers in high-throughput biomedical science to easily plan their experiments online and to collect all related metadata in a database system. The focus is on full factorial experimental designs, but the creation of other experiments is supported. Information storage needs to be implemented in a manner that enables and facilitates re-use of data, either to update it, to use it as input for data analysis, or to share it with collaborators. In order to do so, the existing language of experimenters of these fields needs to be used. Interoperability with respect to existing study exchange formats for biomedical science

needs to be provided. Furthermore, where large studies are concerned, suitable functionality to view metadata is required. In the following, we loosely follow the IEEE Recommended Practice for Software Requirements Specifications<sup>171</sup> to specify the requirements in greater detail.

### 3.2.1 Software Interfaces

The web application is a part of qPortal, a Liferay portal that provides the user authentication. To retrieve or create new projects, experiments, and their biological meta-information, the software is required to communicate with the Open Source Biology Information System (openBIS)<sup>172</sup> via an Application Programming Interface (API). The openBIS application server contains the data model used for experimental design and metadata, including controlled vocabularies that are presented to the user. An interface to an administrative database (User DB) shall be used in order to store administrative information about project investigators, project managers, and any other persons involved in a project. Their respective affiliations are stored in this database as well. Since it must be possible to attach small files related to experimental design to a project, the WebDAV protocol should be used to enable uploads to the data store server. Figure 3.1 gives an overview of the software interfaces and the interconnection of related components.



**Figure 3.1:** Connections between different components of qPortal. Edges to the Experimental Design Wizard portlet denote interfaces to software (blue) and protocols (yellow) the portlet needs to implement.

### 3.2.2 Definitions and Terminology

Table 3.1 gives an overview of the definitions used in the following sections. In order to adapt our user interface for different researchers, we translate parts of the openBIS terminology into more intuitive terms for the use in qPortal. Table 3.2 lists the most important changes. In the following, we will use the qPortal terminology when explaining processes from a user perspective.

**Table 3.1:** Definitions of concepts and terms used in the design of data model and user interface.

Term	Definitions
(User) Space	Spaces are used to define access permissions. Projects, experiments, and samples are part of one space.
Project	Projects are used to group related experiments. Many projects can be part of one space.
Experiment	An experiment describes one step of a study. Multiple samples can be part of one experiment.
Sample	A sample describes one entity that is part of an experiment. Those can be the typical samples used in labs, but also includes the sample source: patients, plants, or cell cultures. Samples can be extracted from other samples in an experiment.
Entity	Umbrella term for instances of openBIS objects. There are different entity types that can be used to distinguish different types of samples and experiments.
Property	openBIS properties are defined categories for a specific kind of metadata. Different property types that can be connected to different entity types.
Sample code	Identifier used by openBIS to uniquely identify every sample object.
User DB	A relational database for storing persons, affiliations, and other general information about a project.

**Table 3.2:** Mapping of openBIS terminology to the qPortal user interface.

openBIS Terminology	qPortal Terminology
(User) space	Project (space)
Project	Sub-Project
Biological Entity	Sample Source
Biological Sample	Sample Extracts
Test Sample	Sample Preparations/Analyte (sample)

### 3.2.3 User Characteristics and User Interface

Main users of the web application are researchers involved in high-throughput biomedical science. Since biologists, lab technicians, as well as bioinformaticians, are often involved in studies, implementation needs to take the differences of perspective and knowledge of this user base into account. Dependent on lab-internal functions, some users may have different authority on creating new sub-projects or changing existing metadata. In addition, administrators of the portal must be able to create new projects and give other users access.

When logged into the system, a user must be provided with the options to add sub-projects and experiments, import a sub-project using different formats, or update metadata of existing samples. When adding experiments without the upload functionality, users shall be able to input new data based on the context of an existing sub-project, to complete data that has previously been input or start a new sub-project. They should be guided through this process. The import functionality should provide information about the different formats that can be uploaded. Uploads shall be processed in order to create new sub-projects or add samples and experiments to existing sub-projects. A third user interface must be provided to enable users to upload a metadata spreadsheet in order to update meta information of samples that already exist in the system.

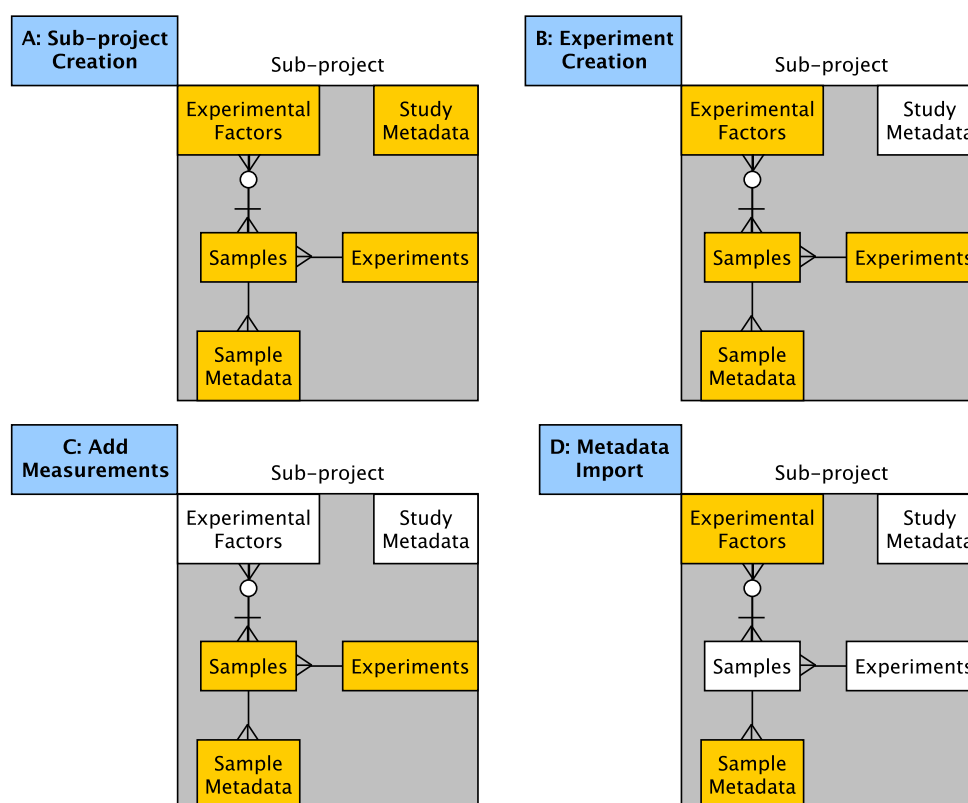
### 3.2.4 Functional Requirements

This section lists the functionality that is needed for users of the portlet interface in more detail. An overview of the main interactions based on the data model is depicted in Figure 3.2.

There are four main use cases that our portlet should facilitate: first, creating a new sub-project with experiments and samples. Second, adding this information in the context of an existing sub-project, for example sample preparations of a second omics layer in a multi-omics study. Third, the creation of new experiments, samples, and their metadata via the upload of a spreadsheet, for example, to import existing experimental designs from external databases. Finally, the editing of existing metadata that has been previously registered, makes up the fourth use case.

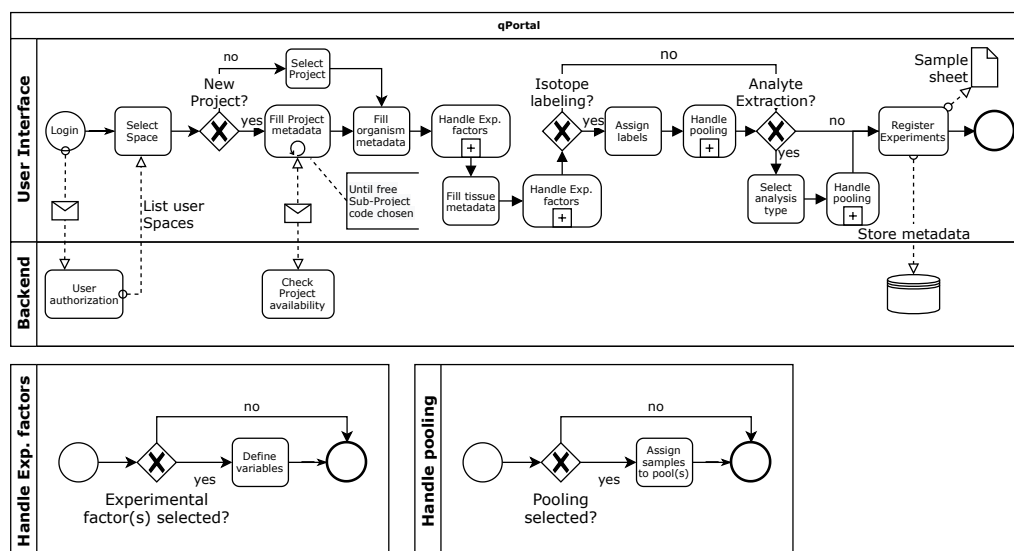
### 3. Modeling of Experimental Designs

---



**Figure 3.2:** Overview of the interplay between the data model and the main functional requirements of the Experimental Design Wizard. The most important parts of a sub-project (openBIS project) and its metadata are displayed. Different functions are colored in blue. Parts of the model that can be updated or created by the respective function are colored in yellow. Other parts are not colored. **A:** Users shall be able to create a complete sub-project containing experimental factors, experiments, samples and different types of metadata. **B:** Users shall be able to add new experiments and samples, including experimental factors, to an existing sub-project. **C:** Users shall be able to add new experiments and samples describing the analyte preparation from existing samples in the system. The experimental design must not be changed by this process. **D:** Users shall be able to upload metadata in order to update experimental factors and other attributes of samples.

Creation of new experiments for new and existing projects via a wizard process represents the two main, and most complex, use cases. The main requirements of these processes are depicted using Business Process Model and Notation (BPMN<sup>173</sup>) in Figure 3.3.



**Figure 3.3:** General features of the user interface and its interplay with the backend for both wizard use cases in BPMN<sup>173</sup>. First, the project context is established using openBIS user authorization, as well as selection or creation of project space and sub-project by the user. Information about study organisms and tissues, as well as optional connected experimental factors are collected. Depending on the type of experiment, isotope labeling or pooling can be specified. After selection of analysis types, the experimental design and related metadata is registered in the openBIS database and a sample sheet with unique sample codes can be downloaded.

As the system should allow work on many different projects with potentially hundreds or thousands of samples, it must be possible to easily find and select the current context of the operations that need to be performed without displaying unrelated information stored in the system. To accomplish this, the project structure (see Table 3.1) can be leveraged. Only after selecting a project space, existing sub-projects of that space are to be presented. Accordingly, users can then either select an existing sub-project to interact with and add new experiments and samples or create a new sub-project, unrelated to existing data.

After establishing the context through project space and sub-project identifiers, it is desirable to collect some administrative metadata for the use case of new sub-project creation. This includes a short descriptive name, a summary of the design, and the selection of different people involved in the new sub-project. A connection to the user database can be leveraged for this functionality.

Once the general information has been provided, users can start with the creation of their experimental design. Data provenance is an important principle of FAIR data management and this includes knowledge about the biological source of measured data. Accordingly, we

### 3. Modeling of Experimental Designs

---

consider it a requirement to collect information of different experimental steps, from patient or model organisms to the sample extract type that is measured using high-throughput methods. For our portlet that means users need to be able to input groups of samples or sample sources for patients (or other organisms), as well as tissue or organism parts. At each of these levels, experimental factors like treatments, genotypes or even different tissues can be part of an experiment and the levels of these factors, as well as the respective number of replicates, must be collected.

Since many modern biomedical experiments use isotope or other labeling methods or pooling for multiplexing reasons, it is also important that users have the option to input information on these steps via the interface. This means assigning labels, for example, different SILAC isotopes, to samples. For pooling, multiple samples need to be collected and assigned a pool name.

The process of designing the experiment commonly ends with the input of different analyte types that are to be measured, such as DNA or proteins extracted from tissues or cell cultures. To support multi-omics experiments, multiple selections of analyte types need to be possible.

Selecting this third level must be optional in order to support experiment types like medical imaging that do not depend on the analysis of small molecules. This option also enables users to complete their experimental design at a later stage of the project. It can also divide the input of experimental metadata into two parts. These are the planning of the experiment on one hand and the preparation and measurement of DNA libraries or protein samples done by lab technicians.

After collection of the experimental design and related metadata, they need to be registered in the data management system. This process leverages the openBIS API. In order to match the stored information to the lab processes, unique identifiers - sample codes - created in openBIS need to be returned to the user. For this purpose, the download of a summary spreadsheet of the experimental design and metadata, including sample codes, needs to be possible.

The third use case, the upload of existing experimental designs via file, is governed by both the respective format, as well as the data model of the system. This means that uploaded metadata needs to be verified with respect to the format specifications, as well as information necessary to complete the experimental design for registration in openBIS. To enable users to successfully import the respective format, the different format options need to be presented to them along with examples or documentation. Verification for openBIS requires the verification of properties that use controlled vocabularies, as well as any necessary sample or organism identifiers. As with the wizard process, the project context needs to be established by selecting project space and sub-project. Displaying summarized information about the uploaded experiments and samples helps scientists verify their experimental design was imported and interpreted correctly. Analogous to the previous use cases, after registration of the experimental design, the download of metadata with unique sample codes is necessary to map experimental

design information to the information stored in openBIS. In this case, the uploaded formats can be used to directly add these identifiers.

These unique identifiers are crucial for the fourth use case, as well. To simplify updating metadata, and avoid making users get accustomed to yet another import format, the only requirement for importing metadata are sample codes of existing experiments. These need to be uploaded in a spreadsheet containing any of the metadata portlet users wish to add or update. This design choice only includes mapping to the sample entities in openBIS, but not to the different properties of the data model. Therefore, this information must be collected following the upload. We choose to do this by displaying the information using the familiar table format given by the upload. The interface can be used to present users with all available options of property types respective to the type of sample codes provided. For example, uploads of codes belonging to the sample source level, will only allow mapping to sample source properties, like taxonomic information, while different property types will be available for DNA sample codes. After users have selected property types of all of their metadata columns, a check for collisions with existing metadata needs to be performed, before users can add or update the metadata of their experiment. In this respect, the different user roles openBIS provides, can be leveraged. While some users may be allowed to overwrite existing information, it is often in the best interest of project management to limit most users to only add missing information.

### **3.2.5 Performance Requirements**

In order to evaluate our implementation further, we propose a number of performance requirements that need to be fulfilled by our web application.

#### **Intuitiveness**

Since multiple different groups of researchers with various backgrounds are involved in modern biomedical projects, the experimental design and metadata collection of a project our implementation must ensure that each of these user groups can effectively work with our web portlet. This means that the terminology used is shared by users or otherwise sufficiently explained. In cases where a specific functionality is performed, the user interfaces for these steps must be easily reachable without detailed knowledge about other, perhaps unknown, procedures.

#### **Response Time**

There are typically three guiding ranges regarding the response time of an interface to a user's actions<sup>174</sup>. A response time of 100 ms or fewer guarantees, that users feel they are directly manipulating objects of the user interface<sup>175</sup>. Longer response times of up to one second make users feel the computer is working but does not let them lose the flow in performing their

task<sup>176</sup>. Functions with response times in the range of 10 s need to be made clear to the user to keep their attention. Any functionality that requires longer run times should be clarified to the user, so they can switch tasks in the meantime. This results in the necessity to re-orient users to the task at hand once the UI is ready.

Based on this information, our goal is to have response times of 100 ms of the user interface for all common processes, one second for intermediary, clearly denoted steps handling many entities, and response times of no more than 10 s for portlet startup, import, and the registration of projects, experiments, and samples.

#### **Scalability**

The size of high-throughput biomedical projects requires that our web portlet must be able to provide its functionality for experiments with hundreds of samples and other entities. This does not only pertain to response time but means that manual information input should be minimized. It also requires that displayed information about experiments is clarified by summarizing it, sorting it, or visualizing it in a helpful manner.

#### **3.2.6 Software System Attributes**

Our portlet needs to fulfill a number of attributes that are independent of the functional requirements of our user interface.

#### **Security**

The application needs to be secure with respect to all use cases, especially due to sensitive experimental metadata that is sometimes entered and stored in the connected systems. First of all, access to the system by unauthorized persons, either involved in different projects or without credentials for the system, must be restricted. This pertains to both the experimental design and sample information, as well as to administrative metadata like involved groups and people. Secondly, users with access to the system must not be able to delete or overwrite metadata or attached data. Power users, while able to overwrite metadata, must be made aware of the consequences if they choose that option.

#### **Maintainability**

The growing diversity of experimental approaches entails the need to make our application easily extensible in response to a growing data model and different high-throughput technologies. In order to facilitate extensions of the existing functionality, the code should be well-structured and openly available (open source) to the community.

## Portability

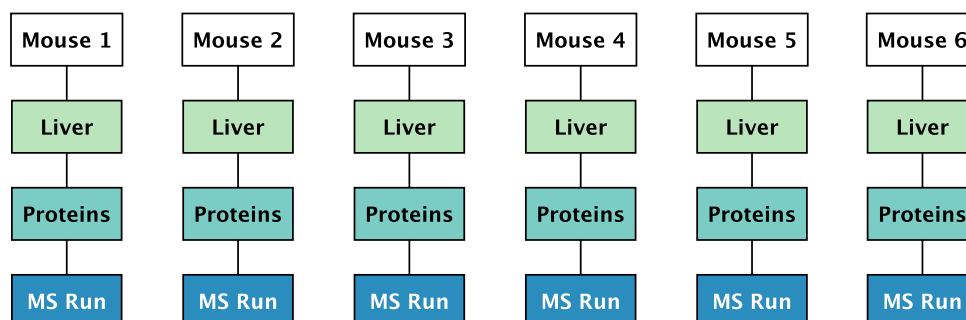
For the purpose of facilitating the software's experimental design options and data sharing capabilities to a significant number of researchers, it needs to be portable to other locations. This includes a comprehensible versioning system as well as options for customization of different instances.

## 3.3 Design and Implementation

### 3.3.1 Backend and Data Model

The Experimental Design Wizard uses openBIS to store metadata. This data management system offers mechanisms for storage and management of raw data and its annotations as well as functionality for managing data access. The software package comprises a raw data store, metadata management via a PostgreSQL database as well as an application server for browsing and managing data and metadata. Data models build the basis for handling big data efficiently. The general openBIS data model comprises five distinct hierarchically ordered levels. Access rights are managed on the top level (spaces), which can contain multiple projects, experiments, samples, and datasets. Custom openBIS data models can be defined by creating specific types for experiments, samples, and datasets. Entities of those types can contain multiple user-defined properties. The structured storage of metadata is an essential component of the system; metadata is attached to the representations of both the concrete samples as well as their intangible experiments.

Experiments measuring biological data typically involve multiple steps, ranging from subject recruitment, treatment, sample extraction to library preparation. Each of these processes and process-related entities like cell cultures, samples and analyte preparations is typically associated with its own set of unique metadata. In order to keep track of all entities and the data generated from them, we use the openBIS data model to define a multi-tier experiment graph: the first tier describes patients, model organisms or similar sample sources with associated metadata like the NCBI Taxonomy<sup>177</sup> identifier. The next tier describes sample extraction of cells, tissues or other material from the aforementioned sample sources. A third tier describing analyte extraction (e.g., of DNA or proteins), can be attached to sample extractions. On this level, we collect metadata pertaining to preparation of the samples for actual data collection on NGS platforms or for mass spectrometry. Since some studies split samples into aliquots or pool them, for example in isotope labeling experiments, both the tissue and the analyte levels can be multi-tier steps in our model. One example of the dependency graph representation of this hierarchical model can be seen in Figure 3.4. Each experimental entity is identifiable by a unique sample or experiment code.



**Figure 3.4:** Simple sample hierarchy graph of a study involving six mice. Liver tissue extraction was performed, protein samples prepared, and measurements taken. Not shown are the various properties of each sample.

For easier retrieval and update of experimental designs, we designed an XML schema definition (XSD) to store all design information related to a single study. Each experimental factor is listed, differentiating between nominal and continuous factors. In the latter case, the type of unit is stored. Factor levels, which include a list of entity identifiers that are related to this factor level, complete the schema.

In addition, the XML format can store information about technology types used for data generation in the study, specific experimental design decisions, as well as additional, user-defined properties of samples that are not covered by the openBIS data model. The XSD allows for validation of input data as well as quick updates.



### 3. Modeling of Experimental Designs

---

**Table 3.3:** Mapping of the ISA-Tab data model to our data model. Some additional information is encapsulated in the mentioned objects as seen in Figure 3.5. For example, the source organism is found in the study file and encapsulated in the Biological Entity created in openBIS.

ISA Data Model	qPortal Data Model	Comments
<b>Investigation</b>	(Space)	Studies can be imported anywhere
Publication	-	
Person/Organisation	-	
<b>Study</b>	Project	
Title	Project name	optional
Description	Project description	optional
Protocol	Long description	optional
Source	Biological Entity	used for species and tissue mapping
Study Design	Exp. design XML	
Factor	Exp. factor (XML)	Hierarchy level of factor is added
Hardware/Software	-	
<b>Assay</b>	Sample Preparation	
Sample	Biological Sample	Used to connect sources to extracts
Extract	Test Sample	Used to map qPortal identifiers to file names
Measurement	Sample Type	Measurement endpoint is translated
Technology	Exp. design XML	Added to Technology section of XML

openBIS entities are created for every unique sample source found in the study. They are defined by the ISA-Tab source name and organism. If included in the specific file, mapping of the NCBI taxon is trivial, since our model uses the same ontology. If this information is missing, this mapping is delegated to the user interface, as shown in Section 3.4.4. The same process is used if the organism part is missing or does not fit existing information in our vocabularies. Similar to the source level, entities for biological samples are created from unique sample names and organism parts found in the file. Hierarchical connections are created between every source and sample whose names are found in the same row of the study file.

Similarly, each assay file carries information about samples and extracts, enabling the creation of analyte samples and a mapping between the lower parts of the hierarchy. The measurement endpoint defined for each assay is mapped to the respective analyte type. For ISA studies with multiple assay files, as they are commonly encountered in multi-omics projects, multiple different analyte samples are created for each tissue sample.

Experimental factors can be part of both study and assay files. Since there is no explicit information in ISA-Tab about which hierarchy level a factor belongs to, we implicitly parse this information from the context. In order to do this, we check the consistency of factor levels with entity identifiers (source and sample names) beginning from the highest hierarchy level. In a study file, the same source ID cannot be related to two different factor levels, but multiple

samples of this entity can have received different treatments each. The same is true for the samples and extracts found in assay files. The ISA model also stores information about the study design. This information is added to our experimental design XML.

### 3.3.3 Software Interfaces and User Management

The Experimental Design Wizard as well as study import and metadata update are implemented as a Java portlet integrated into qPortal. This science gateway runs on top of a Liferay 6.2<sup>179</sup> portal instance using Tomcat 7<sup>180</sup>. The portlets are written using Java 8 and the open-source framework VAADIN 7<sup>159</sup>, which is based on Ajax and Google Web Toolkit.

In the qPortal system, registered users are stored in an in-house Lightweight Directory Access Protocol (LDAP) server connected to openBIS and Liferay. Therefore, we are using the advantages of a single sign-on (SSO) based solution as already implemented for other Grid web applications and portals<sup>181,182</sup>. The resource containing user information can be easily replaced by any comparable protocol (e.g., Crowd<sup>183</sup>) compatible with openBIS and Liferay. Access to data and metadata is regulated on different levels. The primary login mechanism is placed on the Liferay landing page, followed by a delegation mechanism to the back-end database. Each user or defined user group can be assigned multiple roles, regulating access to openBIS spaces that might include several projects and the corresponding data. Therefore users are only able to access data connected to projects in spaces to which they have been granted access.

### 3.3.4 Guided Metadata Management

In order to simplify the work with large studies, we implement several methods to help with the input and integration of metadata from different sources. As described in Section 3.2.4, we enable users to upload a spreadsheet containing existing sample codes and any additional metadata. Our portlet automatically queries openBIS to determine the entity type of the related samples and returns a list of possible metadata properties. Users can select one of these types for each of the column names in their document and update the selected metadata in openBIS. After selecting a property type, different types of verification are performed in order to guide the user. Data in the respective column is checked for consistency with the selected data type. Inconsistent data can not be updated. Column data is also compared to the metadata already present in the data management system for the specified sample codes.

Users can also select complex properties like experimental factors. In this case, additional information like a name for the factor is needed. The number and names of factor levels can be created from the provided cells in the selected column.

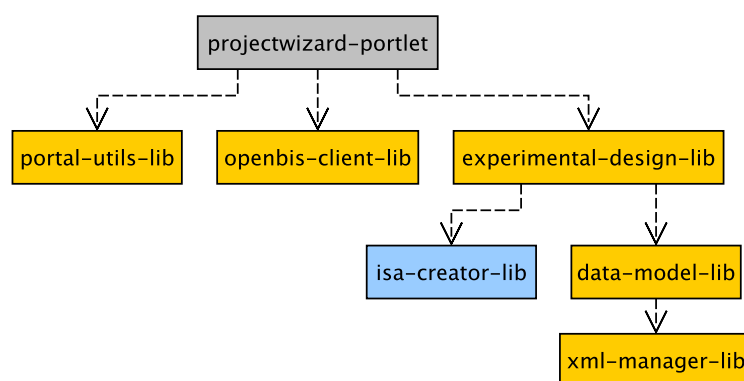
Some of the property types in openBIS are complex types that follow controlled vocabularies containing keys, which often are ontological identifiers, and a respective humanly-readable

label. When selecting complex property types that use controlled vocabularies, we search for similar labels in the respective vocabulary. If no label is found, users are provided with the option to map the unknown metadata to the available, valid values from the vocabulary.

This functionality is also used for the import of new sub-projects. For example, if species, tissues or analyte types are not found in the respective openBIS vocabularies, users are able to select the correct property values corresponding to their respective inputs.

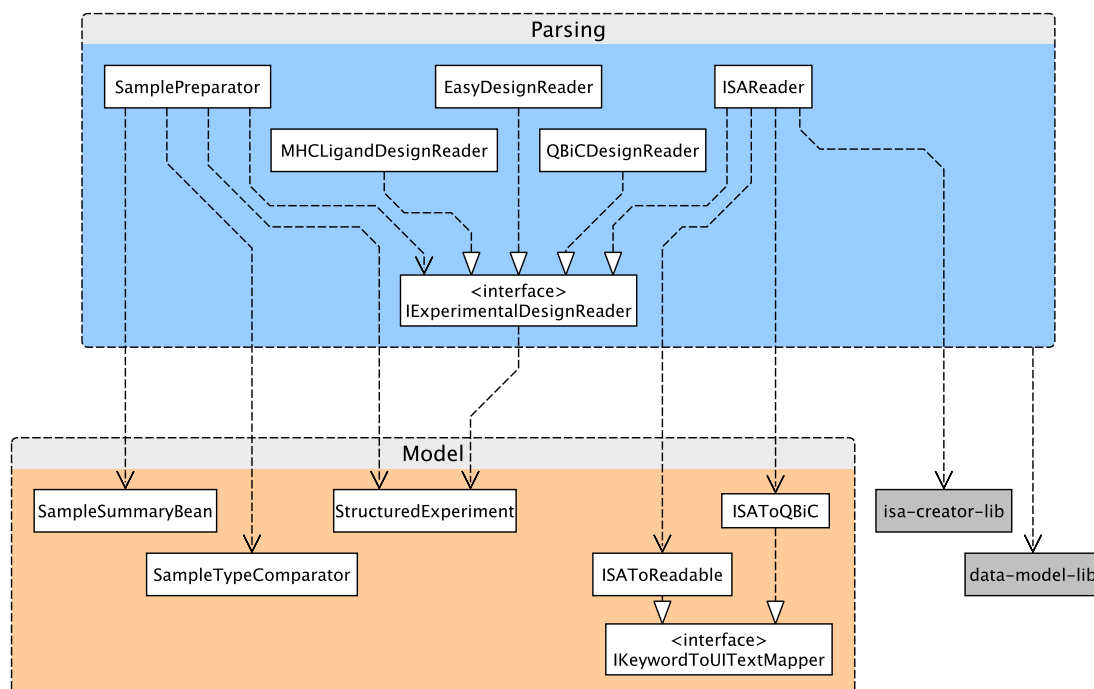
#### 3.3.5 Software Components

In order to facilitate maintainability and portability of our software, the Experimental Design Wizard is modularized into several libraries as shown in Figure 3.6.



**Figure 3.6:** Dependencies of the Experimental Design Wizard and its libraries. Grey: wizard portlet itself. Yellow: libraries used by the portlet. Light blue: third-party libraries that have been adapted for our use. Third party libraries whose code has not been changed are not shown.

Functionality that is used by all qPortal portlets is organized in the portal-utils library. This includes functions that communicate with the Liferay environment to get the username and similar information. The openbis-client library provides convenient methods to communicate with the openBIS API in order to fetch information from the data management system or register new projects, experiments and samples. The core library of the Experimental Design Wizard is experimental-design-lib. As shown in Figure 3.7 it implements parsers for different reporting formats that check imports for consistency.



**Figure 3.7:** UML class diagram showing the relationships between different classes and dependencies of the experimental-design library used by the Experimental Design Wizard. The library is split into the main package responsible for parsing different experimental design formats and a model package used to produce entities for openBIS and other applications. Grey: other libraries used by experimental-design-lib classes.

They provide a translation into the openBIS data model, a summary of species and sample types and the experiment aggregation graph described in the following chapter. Interfaces are used to make extensions for other formats or data models easier. The library itself depends on the data-model library, which describes the sample and experiment model and provides convenience functions for different identifiers and sample sorting. Factors, levels and units of the experimental design are stored as XML for fast retrieval and update. These functions are implemented in xml-manager-lib. In order to parse the ISA-Tab standard as described in 3.3.2, isa-creator-lib was created from the ISAcreeator application<sup>17</sup>. All of our libraries, as well as several third-party libraries, are included in the Experimental Design Wizard using the build automation tool Maven<sup>184</sup>.

The main portlet is structured using the model-view-controller design pattern. Since most of the openBIS-related data model is accessed via the data-model library, the model package predominantly includes classes to store complex experimental input related to mass spectrometry or ligandomics. The view includes the different steps of the Experimental Design Wizard. Controllers handle both the project import and project creation via wizard steps. An interface is used since in both cases similar steps have to be performed. Since the creation of

entities and metadata in the openBIS system is one of the main tasks required of the portlet, the functionality of openbis-client-lib is extended in the metadata registration package of the Experimental Design Wizard. Here, the data collected via project import or wizard steps are split into parts that can be registered in the system consecutively using a background thread. In this process a comparison with existing projects, experiments and samples is performed. To inform users of the progress and status, a progress bar and helpful information are displayed.

#### 3.3.6 Data Transfer

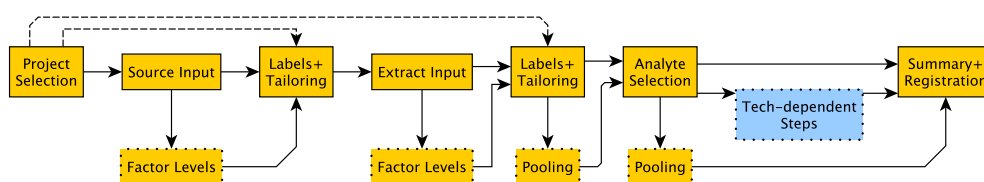
Data transfer from different locations to the central data repository is realized using rsync, as implemented in the openBIS Datamover<sup>172,185</sup>. Data is synced from the source, for example, the genome sequencer to defined folders on remote storage. Identifiers created using the Experimental Design Wizard are used to tag the files that have been measured for each individual sample. Checksums allow monitoring of data integrity while transferring data to the final catalog via openBIS dropboxes according to their source and data type. Every dropbox implements an Extract Transform Load (ETL) routine. These scripts are based on Jython and handle the raw data and possibly connected metadata files. Depending on the needed ETL process, additional external scripts can be called. In some cases, additional information, that has not been registered via the Experimental Design Wizard, is collected when actually preparing the samples and measuring the data. ETL scripts complete the annotation task by creating new data model entities like experiments, samples and datasets and finalize the experimental model by extracting metadata from incoming files. This information is then stored as content of defined properties of the new entities. The entities are then connected to existing entities in the database through identifiers. The raw data and any files are files converted via the ETL script are then connected to datasets and moved to the data store.

### 3.4 Results

The different functional requirements of the Experimental Design Wizard are realized by different tabs in the user interface of the portlet. An administrative tab provides administrators with the functionality to add new openBIS spaces for specific users. Details of the remaining tabs of the portlet are explained in the following sections.

#### 3.4.1 Interactive Study Design

The core part of the Experimental Design Wizard is sample creation via different steps that can be seen in Figure 3.8. In order to start this process, users select the project (corresponds to a space in openBIS, see Section 3.2) and sub-project under which new experimental steps and samples will be stored.



**Figure 3.8:** Flow diagram of the different steps of the Experimental Design Wizard. Dotted nodes indicate the optional steps sample pooling and factor level input. Dashed lines denote transitions that enable users to skip to existing designs in order to add new extracts to sample sources or prepare new analyte samples from extracts. Technology-specific steps as they exist for proteomics are denoted by the blue node.

Here, a short name and a description can be added and the persons involved in the project can be selected from dropdown menus. For existing sub-projects, the stored sample hierarchy can be used to derive new sample extracts or sample preparations in the following steps.

Figure 3.9 shows the second step of the wizard process. Here, new sample sources can be input into the system.

The screenshot shows the 'Sample Sources' step of the wizard. At the top, there are three tabs: 'Create Project', 'Import Project', and 'Update Metadata'. Below these are four progress indicators: '1. Project', '2. Sample Sources' (which is highlighted), '3. Biol. Variables', and '4. Summary'. The main content area is titled 'Sample Sources' with a help icon. It contains several input fields: 'Experimental Step Name' (a text box), 'Experimental Variable' (two dropdown menus, one set to 'Phenotype' and the other to 'Genotype'), 'Species' (a dropdown menu set to 'Mus musculus'), 'Contact Person' (a dropdown menu with a refresh icon), and 'How many identical biological replicates (e.g. animals) per group are' (a text box set to '1').

**Figure 3.9:** Sample source selection step of the Experimental Design Wizard. Users have to select a species and can add optional experimental factors describing their experiment. Here, the two factors genotype and phenotype have been added.

Users select the studied species from a list taken from the NCBI taxonomy ontology. They can add multiple common experimental factors from a pre-defined list or define their own factor name. If there are biological replicates, this can be specified at the bottom of this step. Selecting one or more factors adds the factor levels step to the wizard.

### 3. Modeling of Experimental Designs

As seen in Figure 3.10, users can input multiple levels for each of the factor labels they have selected in the previous step.

Figure 3.10 shows two screenshots of the experimental design interface. The left screenshot displays the 'Phenotype' tab with 'affected' and 'control' entered in the 'Values' field. The right screenshot displays the 'Genotype' tab with 'WT' and 'KO' entered in the 'Values' field. Below the 'Values' field, there are radio buttons for 'Continuous' and 'Categorical', with 'Categorical' selected. A 'Preview of Combinations' table is shown, listing the combinations of factors and the number of samples for each combination.

Factors	Samples
affected ; WT	1
affected ; KO	1
control ; WT	1
control ; KO	1

**Figure 3.10: Left:** Tabs are created from the previously selected factors. A user has added two factor levels for the categorical factor *phenotype*. **Right:** Once the levels for all factors are completed, a preview of all study group combinations and the number of related samples is created based on the previously selected number of replicates. Users can adjust the amount of samples in each group, if their experimental design is unbalanced.

From this information and the chosen replicates, a full factorial design is created. For every study group, the number of entities - sources or samples - is displayed in a table. Users can fine-tune their design, by changing the default value, before the actual entities are created. In the following tailoring step seen in Figure 3.11, these entities are listed in a table.

Sample Sources Tailoring [?](#)

6 Samples

Secondary Name	External DB ID	phenotype	genotype	Customize
affected ; WT	mouse 1	affected	WT	
affected ; KO	mouse 2	affected	KO	
control ; WT	mouse 3	control	WT	
control ; WT	mouse 4	control	WT	
control ; KO	mouse 5	control	KO	
control ; KO	mouse 6	control	KO	

**Figure 3.11:** Sample sources are summarized in the tailoring step. Names and identifiers can be changed by the user.

They are automatically named by their factor levels if users choose not to rename them. Afterward, users follow similar steps for the creation of the next hierarchy level, sample extracts. Tissues or species themselves can be chosen as an experimental factor for comparative studies. Subsequently, users select the number of different tissues in their study and can then pick multiple entries from the controlled vocabulary in question to denote the factor levels, as can be seen in Figure 3.12.

Tissues

Tissues 1 \*

Whole Blood

Tissues 2 \*

Lymph

Preview of Combinations

Factors	Samples
(2) affected ; WT ; Whole Blood	1
(2) affected ; WT ; Lymph	1
(3) affected ; KO ; Whole Blood	1
(3) affected ; KO ; Lymph	1
(4) control ; WT ; Whole Blood	1
(4) control ; WT ; Lymph	1
(5) control ; WT ; Whole Blood	1
(5) control ; WT ; Lymph	1
(6) control ; KO ; Whole Blood	1
(6) control ; KO ; Lymph	1
(7) control ; KO ; Whole Blood	1
(7) control ; KO ; Lymph	1

**Figure 3.12:** Since tissue was selected as experimental factor, the following factor level step allows users to select the respective tissues from the tissue vocabulary. Resulting study groups are created by including the experimental design of the sample source level.

The resulting factor level combinations of the extract tier are combined with any existing levels of the sample source tier, creating the final full factorial experimental design including all specified factors of both hierarchy levels. The most notable difference to the sample source step is that users can select sample pooling and isotope labeling starting with the extract level. Labeling type and the sample labels themselves are selected in the respective tailoring step.

### 3. Modeling of Experimental Designs

Pooling presents the user with a table of their previously created and optionally labeled samples as seen in Figure 3.13.

**Pool 1**

Secondary Name  
Sample pool

Sample Pool

ID	Secondary Name	Lab ID	Undo
2	sample 2	R02	↺
1	sample 1	R01	↺

**Unpooled Samples** | Pooled Samples

	Secondary Name	Lab ID	silac
3	sample 3	R03	SILAC medium

**Figure 3.13:** Pooling step of the Experimental Design Wizard. Available samples are displayed on the right side. A user has added two SILAC-labeled samples to a new pool on the left. For every pool, one new sample referencing its respective pooled samples is created in the next step.

Drag and drop can now be used to place arbitrary groups of samples into a pooling group. Both samples and completed pools are then transferred to the following step.

In the analysis method step, users can select if and how analyte samples are to be prepared from their sample extracts. For each selected analyte type like DNA or lipids, a sample is attached to each sample extract. Additional pooling can be performed for these analytes. Dependent on the type of analyte, technology-specific steps can follow as described using the example of proteomics and peptidomics in Section 3.4.5. Otherwise, the resulting entities are shown in the summary step, as seen in Figure 3.14.

Sample Registration

Summary

Type	Content	Samples
Sample Sources	Mus Musculus	6
Split Sample Extracts	Lymph, Whole Blood	12
Sample Preparations	RNA	12

Download Spreadsheet

Register All Samples

**Figure 3.14:** Summary and registration step of the Experimental Design Wizard. Organism, tissues and analytes are summarized with their respective numbers of entities. Users can start the registration of their sub-project by clicking on the respective button.

Here, the type of organism, tissue, and measured analytes are highlighted and users can see the number of entities for each hierarchy level. Data registration can then be started with the click of a button. Following the registration, the last step allows users to download different spreadsheets with barcodes and metadata for each hierarchy level. Small attachments pertaining to the experimental design of the study can be uploaded. A link allows access to the newly registered project via the project browser portlet of qPortal, as described by Mohr et al.<sup>186</sup>.

### 3.4.2 Response Time

We collected data for several use-case scenarios (Tab. 3.4) to evaluate response times of the Experimental Design Wizard.

**Table 3.4:** Response times of the Experimental Design Wizard portlet GUI for different use cases, grouped by test cases that were performed together. Unless otherwise specified, response times denote the time after specified information has been filled in, the *next* button has been clicked, until the UI in the next step has finished loading and is responsive.

Process	Response time [ms]
First loading of portlet	2,448
Context displayed after selecting existing sub-project	542
Adding 20 sample extracts to existing sources	233
Selecting 10 replicates for new sample sources	342
Adding 10 extracts (one per source)	317
Adding RNA and DNA samples per extract, showing summary	300
Selecting 100 replicates for new sample sources	583
Adding 100 extracts (one per source)	633
Selecting 1,000 replicates for new sample sources	3,150
Selecting 2 experimental factors and 50 source replicates	383
Creating a $2^2$ design from the factors, resulting in 200 sources	995
Selecting tissue as an experimental factor	183
Selecting two different tissues types, resulting in 400 extracts	1,800
Adding RNA and DNA samples per extract, showing summary	1,567
Importing ISA-Tab study containing 2,160 entities	389

Aside from the first loading of the portlet (2.4 s), which involves the collection of all openBIS projects and vocabularies, our benchmark shows that interactions with large amounts of objects lead to the highest delays in the loading of different pages or UI elements. Specifically, creating 1,000 sample source, as well as 400 sample extract objects in a large, multifactorial experiment, and displaying them in tables, took 3.2 s and 1.8 s, respectively. Further tests reveal that the largest part of these delays can be allotted to the building of table elements, themselves, while

### 3. Modeling of Experimental Designs

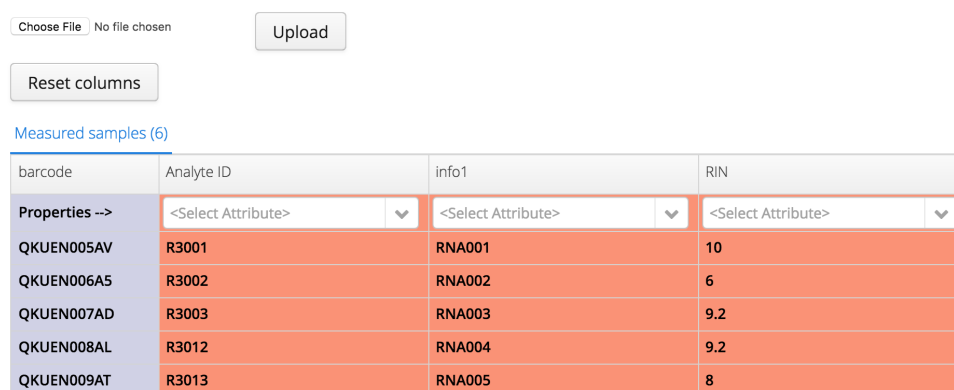
creation, linking, and annotation of the sample source and extract objects take a much shorter amount of time.

In general, the number of replicates and complexity of experimental design do not lead to increased loading times of the user interface for the tested experiment sizes. We created a full-factorial design using two different genotypes, two phenotypes with 50 biological replicates for each study group. We used two different tissue extracts per each of these 200 sample sources and selected the preparation of RNA and DNA from each tissue extract, resulting in 1,400 entities. While the response times of the involved steps varied between 0.2 s and 1.8 s, the slower processes were limited to the clearly denoted transition between the wizard's steps. For interactive processes in each step, i.e. displaying project context or adding factor levels, the UI remained fast and responsive.

One alternative to creating studies with such large sample sizes via the wizard process, is the import via metadata formats. The time between the import of an ISA-Tab study containing 2,160 entities (E-GEOD-3210 from the Personalized NSAID Therapeutics Consortium (PENTACON, [www.pentacohq.org](http://www.pentacohq.org))) and the notification of successful import was 0.4 s.

#### 3.4.3 Metadata Management

In order to easily update stored metadata or add new annotations, any tab-delimited file containing registered entity identifiers and related meta information about samples can be uploaded. The interface can be seen in Figure 3.15.



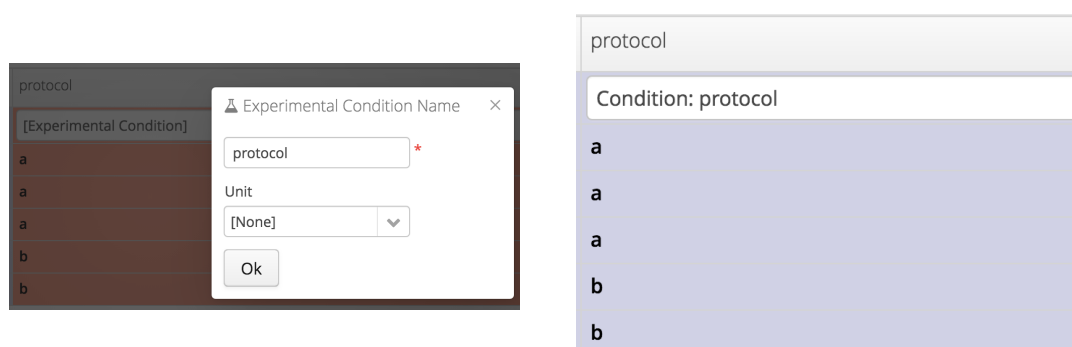
barcode	Analyte ID	info1	RIN
QKUN005AV	R3001	RNA001	10
QKUN006A5	R3002	RNA002	6
QKUN007AD	R3003	RNA003	9.2
QKUN008AL	R3012	RNA004	9.2
QKUN009AT	R3013	RNA005	8

**Figure 3.15:** View of the metadata update tab after uploading a spreadsheet with sample information. At the top, users can select the property type fitting the metadata of a column they want to store. Identifiers of registered samples (left column) are automatically found in the system. For other columns, all data is listed. Red columns denote missing user input.

It allows to either discard or map user-defined column labels to property types from the openBIS data model. This selection is context-specific: metadata of sample source identifiers can only be mapped to sample source-related properties, such as the organism, while metadata

of RNA samples may be mapped to information like the RNA Integrity Number (RIN). When selecting properties that are restricted to a specific format, metadata is checked for consistency with these types and users receive feedback, if those restrictions are not met. For example, only whole numbers can be stored for a property type that is measured in integers.

For controlled vocabularies as they are used for tissues or the NCBI taxonomy, users can map unknown metadata values to the vocabularies through a dialog window. Here, valid vocabulary values are presented in a searchable dropdown menu. Users only need to perform this mapping once, even if multiple samples or sample sources are to be updated with the value in question, limiting effort for large studies. The portlet also allows updates of complex properties outside of the openBIS data model. Any custom property consisting of name, value and an optional unit can be added to different entities. The experimental design of the study can be updated as well. If a factor already exists in this experimental design, samples, levels or both are updated. If the specified factor is new to this experimental design, it is added to the respective data structure. The dialog option showing this use case can be seen in Figure 3.16.



**Figure 3.16:** **Left:** Complex properties like experimental factors can be added via the metadata upload. After selecting the respective property, a pop-up window enables users to choose a factor name and unit. **Right:** After a factor has been specified — in this case *protocol* — the system automatically creates the factor levels from the uploaded data. Levels for this factor are *a* and *b*. As a result of a successful property selection without collisions, the column changes to a blue color.

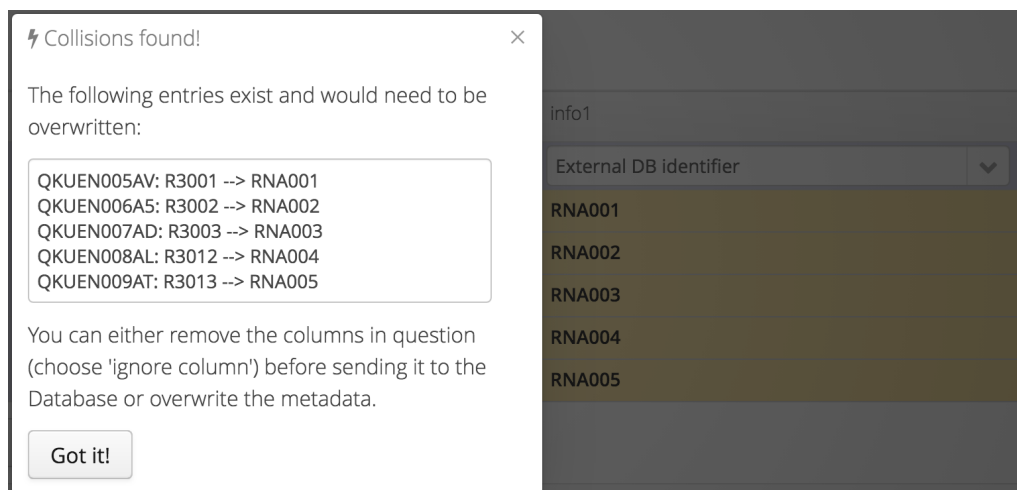
Once users have mapped or discarded all unknown column names of their metadata to the underlying property model, a check for needed annotation updates is performed. Existing data is ignored while missing data is marked for update. If at least one property value of at least one existing entity would be changed by an update, users are informed of all of these data collisions. This pertains also to experimental factors, so the experimental design can not be changed by accident.

All changes to the annotations can then be committed via the press of a button.

We have implemented the process to be able to utilise pre-defined user privileges. Dependent on the defined roles, some users may only be able to add new information while changing

### 3. Modeling of Experimental Designs

existing metadata can be reserved to principal investigators or similar project leaders. An example of the warning dialog shown when metadata is about to be overwritten can be seen in Figure 3.17.



**Figure 3.17:** After selecting a property for each column to be registered in the system, collisions with existing values for these samples are displayed to the user.

#### 3.4.4 Study Import and Interoperability

We provide a number of different import options in order to make it easier for researchers to work with large projects and save time with the input of existing studies and their metadata. Users can import an openBIS-based format that allows specification of project, space and experiments. This is primarily used to create projects after users submit a draft using the wizard process. We also provide a spreadsheet format that allows study creation by specifying a set of minimum information that is often used in biomedical projects. For each sample, an identifying name, tissue type, species and sample source identifier has to be provided, in order to identify biological replicates. If analyte samples are to be created, users also have to provide their IDs and measurement type. Experimental factors and their levels can be added in a new column. Most notably, we allow the import of ISA-Tab via our graphical user interface.

Users can preview the aggregated sample graph (as described in Chapter 4) and need to select the project and sub-project into which the study should be imported. If a new sub-project is selected, the study name and description are automatically added to the respective fields of the user interface. Organism, tissue and measurement type are parsed from the uploaded files and compared to controlled vocabularies. Unique missing entries are marked red and prompt the user to choose a suitable alternative from the provided list. The graphical user interface can be seen in Figure 3.18.

Study  
BII-S-2

Preview Sample Graph

Project

Sub-Project

New Sub-Project

Sample information (please complete)

RNA

unspecified organ \*

Saccharomyces cerevisiae

**Figure 3.18:** User interface for importing of ISA-Tab studies. Any study contained in the ISA-Tab upload can be selected in order to register it in the system. Mandatory experiment metadata is mapped to the openBIS data model and missing information is highlighted. Here, an unspecified tissue could not be found in a controlled vocabulary and the user can select an alternative via a drop-down.

Once the form has been completed with all the information, the summary table is shown (Fig. 3.19) and users can register the study in the system.

Project  
CONFERENCE\_DEMO

Sub-Project

New Sub-Project  
QGMVK

Short name  
A time course analysis o

Sample information (please complete)

RNA

unspecified organ \*

unknown tissue

Saccharomyces cerevisiae

Saccharomyces cerevisiae

Description \*

Comprehensive high-throughput analyses at the levels of mRNAs, proteins, and metabolites, and studies on gene expression patterns are required for systems

Summary

Type	Content	Samples
Sample Sources	Saccharomyces Cerevisiae	2
Split Sample Extracts	Unknown Tissue	14
Sample Preparations	RNA	14

**Figure 3.19:** Completion of missing information allows users to register the imported study. **Left:** Selecting a new project automatically inserts the ISA study title as project (short) name. **Right:** Likewise, the ISA study description is automatically added to the description field. A summary of the imported samples is shown.

### 3.4.5 Extensibility

The hierarchical approach of our data model is easily extensible by adding new tiers of samples. Table 3.5 shows the number of experimental studies and samples created using the Experimental Design Wizard interface.

### 3. Modeling of Experimental Designs

---

**Table 3.5:** Number of projects and samples (only analyte level) registered using the Experimental Design Wizard interface of qPortal as of 03.10.2021. Data is sorted by omics technologies used. Projects using multiple omics technologies are only counted once in the summary.

Technology	Number of projects	Number of measured samples
Microarrays	18	701
NGS	279	19,010
Proteomics	364	19,482
Ligandomics	17	1,739
Imaging	2	172
Summary	651	41,104

#### Proteomics Studies

Our implementation of a proteomics module shows the adaptability of our data model as well as the wizard approach. We created two additional steps to empower researchers to further annotate their proteomics and peptidomics projects. Selecting the *proteins* entry in the analysis method step enables users to select different proteomics options like protein purification. This option also adds a proteomics step to the wizard. If the option *measure peptides* is selected, an additional peptidomics step follows. Both steps enable users to annotate the splitting of previously created samples into fractions using different gels. Enrichment processes like phosphopeptide enrichment of samples can also be modeled. Mass spectrometers, LC-MS methods, and other mass spectrometry options can be selected and wash runs added, for which a reference sample can be specified to indicate the time of the wash run. In the proteomics step, one or more digestion enzymes can be specified per sample. Only for samples that are treated with digestion enzymes, peptide samples are derived for the optional peptidomics step. As with the general wizard process, the summary step concludes study registration. In the data model, sample fractions and digested samples are modeled as analyte samples and attached to the parent sample they are derived from. Information related to mass spectrometry measurements is stored as a new sample type and attached to the respective protein or peptide samples. We explore these types of more complex sample hierarchies in Chapter 4.

#### MHC Ligandomics Studies

The Major Histocompatibility Complexes (MHC) I and II are two protein receptors on the cell surface that are responsible for presenting peptides from intracellular and extracellular proteins to the immune system. The field of MHC ligandomics uses antibodies to capture a specific type of MHC. The bound peptide ligands can be measured by mass spectrometry methods to use them to define the repertoire of peptides presented in healthy or sick individuals. This has implications for personalized medicine and cancer therapies.

We adapted our data model by adding a new experiment type, *MHC Ligand Extraction* and a respective sample type, *MHC Ligand Extract*. After input of the basic experimental design, users can select options to specify the type and mass of antibodies used. The respective ligand extracts of type I or II are automatically created based on the selected antibodies.

To speed up the creation of larger projects, we also created the possibility to upload a complete MHC ligandomics experiment spreadsheet. Here, additional metadata like sample volume or mass, cell types, cancer-related information, the HLA typing of patients and more information about the mass spectrometry measurements can be directly registered into the system. As with our other import approaches, users can download a file that provides a mapping between the identifiers of their data and the newly created samples in our system, simplifying data upload.

### 3.5 Discussion

A central issue of reproducibility is missing metadata annotation<sup>20,24,25</sup>. The creation of web-platforms for scientific research has created new avenues of metadata collection<sup>151–153</sup>. However, the variety of involved fields, as well as the complexity of studies, call for new approaches. As a contribution to this problem, we have developed an Experimental Design Wizard, a web-application embedded into qPortal allowing users to create a full factorial experimental design. In a full factorial design, every independent experimental variable input by the user is multiplied by the number of replicates, creating the same amount of samples for each permutation of factor levels. In reality, full factorial designs are rare, so our application enables users to remove unnecessary samples as well as specify a higher number of replicates for single factors.

Thanks to the modularity of our tool, the Experimental Design Wizard can be easily adapted for more complex and technology-specific metadata collection. Extensive functionality for supporting proteomics and ligandomics experiments has been added.

The approach of starting extensive metadata collection before the experiment is carried out has numerous advantages. First, time and money can be saved, because the study design allows for the estimation of statistical power before experiments are performed. Secondly, mistakes in study design or sample handling can be traced back more easily and with higher confidence. These aspects are crucial, as one of the issues of reproducibility has been identified as the lack of good study design.

Our approach to annotate metadata is intuitive, as it allows users to focus on the core question of their studies, by making the experimental design the heart of study creation. Wizards, in general, are a proven software tool to break down a longer process for users<sup>187</sup>, in order not to overburden them with too much information. This also enables context-dependent skipping of steps to get to the relevant information, so researchers don't have to invest time

### 3. Modeling of Experimental Designs

---

in the annotation that does not fit their use case. In the same way, the clear cut between experimental design input and annotation of further steps is tailored to the different groups of users involved in modern high-throughput biomedical projects. Examples of this are steps that allow the collection of proteomics and ligandomics metadata.

We have further shown that, in general, the responsiveness of our application meets common standards for established use cases. The startup time of the portlet was found to take a few seconds. In such cases, users are consistently more accepting of waiting times, as they have not started a process that demands their attention. Additionally, the Vaadin framework automatically displays loading indicators, which can help clarify to users that the computer is working<sup>176</sup>. While response time of GUI elements inside the different steps in the wizard is almost instantaneous, the computations taking place between steps take considerably longer. Nonetheless, the large majority of these loading times are smaller or equal to one second, which is well within the acceptable range<sup>174</sup>. In cases where many hundreds or thousands of entities are listed in the portlet, response time decreases. In practice, these cases rarely arise, as we support large studies by enabling the import of different formats and our metadata update functionality. Our tests have shown that the import of such experimental designs is simple and responsive. Nonetheless, since the bottleneck has been shown to be the display of entities itself, the process could be sped up by splitting the large number of entities in a multi-page table or by using Lazy Loading strategies<sup>188</sup>. The registration of completed sub-projects with experimental designs, experiments and samples via the openBIS API can take a few seconds for very large studies. In these cases, users are informed of the progress in order to keep their attention.

While the general security of qPortal is realized by a delegation of user management to the Liferay system, internal security of the Experimental Design Wizard software is provided by a strict separation of user roles. Only users registered in the openBIS system are allowed to access metadata. Only power users are able to overwrite metadata after being informed of the consequences. While data itself is attached to entities via ETL processes, users of the portlet are not able to delete or edit it.

The Experimental Design Wizard is both modularized into multiple Java libraries managed by Maven and the model-view-control-based packages of the project itself in order to be maintainable. The success of this approach can be seen by the extension of the existing software to include new, extended experiment models for ligandomics as well as new external import formats like ISA-Tab. This approach also leads to easier portability. Adaptations to other data models only need to be made in few libraries and specific libraries can be substituted. The Experimental Design Wizard is being used in different locations as a part of the qPortal package.

Our approach adheres to the FAIR guiding principles for scientific data management and stewardship<sup>27</sup> in the following ways. First, the Experimental Design Wizard enables users to generate large study designs including metadata. If additional metadata needs to be added, fast

methods to update these large designs can be used. Meta information is stored using openBIS, which provides the connection to raw data. openBIS indexes the data and facilitates efficient searching based on our metadata, making the data findable.

The second criterium of the FAIR principles is the accessibility of data. We follow these guidelines by using open and free protocols for data access through our web-portal. Security is delegated to the Liferay platform, using common authentication processes as described in Mohr et al.<sup>186</sup>.

Data must also be interoperable, meaning that metadata must use a formal, accessible, shared, and broadly applicable language. We contribute to this guideline by using controlled vocabularies, such as the NCBI taxonomy database. We also provide interoperability with the ISA-Tab format. Export functionality based on our data model, including to further public databases, like the Gene Expression Omnibus (GEO), is under development.

The last of the FAIR principles is reusability. It is outside of the scope of the Experimental Design Wizard to track data provenance. However, the same models that enable the storage of the biological provenance from patient to measurement can and are used for the normalization and analysis steps that transform raw data to results.



## Chapter 4

# Visual Exploration of Experimental Designs

Some content of this chapter is part of the manuscript:

---

*Interactive Visualization for Large-Scale Multi-Factorial Research Designs*

Andreas Friedrich, Luis de la Garza, Oliver Kohlbacher, Sven Nahnsen *International Conference on Data Integration in the Life Sciences* Springer 75–84 (2018)

---

### 4.1 Introduction

As described in the last chapter, there are a variety of methods to simplify the creation and maintenance of high-throughput biomedical studies. Science gateways and data management systems designed for biological data do not only help with metadata collection. They also present this information to users, often structured by different projects or analysis types<sup>151,153,172</sup> and often focusing on workflows and their annotation<sup>152</sup>. However, in order to facilitate real reproducibility and replicability, it must also become easier to understand and work with this information after the project has been concluded. Especially for projects with many replicates and independent variables, factorial designs can lead to a large number of cases that can already make it difficult for researchers to keep sight of the big picture. These problems are exacerbated when researchers unfamiliar with a study and its design want to work with the data.

Public data repositories like PRIDE<sup>141</sup> or ENA<sup>2</sup> make studies findable for researchers, display collected metadata, and allow their download along with raw and analyzed scientific data. If visualization is available, it is often limited to the context of the repository itself, showing how many datasets or which types of data are stored. Processing of metadata for visualization of single projects or their experiments is generally not the aim of these approaches.

Among standards for experiment sharing, ISA-Tab stands out as an important step to bridge this gap. Different tools have been created to use the experimental information that they collect<sup>17,18</sup>. linkedISA leverages the data provided to create a semantic, interoperable presentation and shows how implicitly defined study groups can be extracted from ISA-Tab. These groups are summarized and listed in Bio-GraphIn, a graph-based repository for biological experimental data<sup>189</sup>. With the growing complexity of biological experiments and especially the communication thereof, efficient visualizations are indispensable. However, most of the previous work has been focused on connecting experiments to ontology frameworks and making it machine-readable. While Bio-GraphIn presents a list of study groups, this type of presentation can become difficult to comprehend for huge experiments involving many experimental factors and other metadata. More information can only be obtained by displaying huge tables of samples. It is therefore not only difficult to grasp the complexity of a study, but also to assess the quality of ISA files created by researchers. While tools can prevent mistakes violating the format standards, clerical errors that lead to the reporting of a faulty design are not obvious.

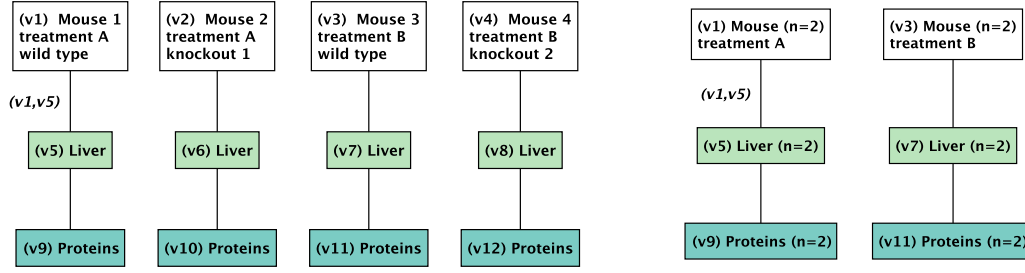
Here, we build on our intuitive interface for experiment creation leveraging proven experimental design concepts like full-factorial study design<sup>190</sup>. To connect experimental designs with data integration, we provide an interactive visualization tool that can aggregate complex study designs based on involved species, tissues, analytes and experimental factors into an intuitive experiment graph. In an effort to comply with existing standards while allowing easy options to manage high-throughput experiments, we provide interoperability with the ISA-Tab format. The highly modular structure makes our tool a good starting point for further developments in the area of quality control and data exploration of biomedical studies.

## 4.2 Materials and Methods

### 4.2.1 Generation of Aggregation Graphs

The complete experiment graph as described in 3.3.1 provides a full overview of all entities involved in a study. We aggregate nodes by identity or similarity of factor levels, tissues or species to arrive at a simplified, more compact version of the experiment graph defined as follows:

Let  $G = (V_G, E_G)$  be a sample graph with vertices  $v \in V_G$  denoting each of these entities in an experiment and edges  $(v, w) \in E_G$  denoting the extraction of entity  $w$  from entity  $v$  in an experimental step as seen on the left side of Figure 4.1.



**Figure 4.1:** **Left:** visual representation of the full experiment graph  $G$  of an experiment on four mice consisting of nodes  $v_1 \dots v_{12}$ . Factor  $f_1$  (treatment) consists of two levels of which each is replicated once. Factor  $f_2$  (genotype) consists of three levels, of which only the wild type is replicated. Inherited factor levels for connected liver and protein extracts are not shown. **Right:** aggregation graph  $H_1$  on  $f_1$ . Factor  $f_2$  is ignored. The two replicates for both treatments are aggregated into single nodes  $v_1$  and  $v_3$ , while the number of samples  $n$  is incremented. Since extracted entities inherit factor levels, they too are aggregated.

Let further  $f_1 \dots f_n$  be a set of experimental factors on a subset of these entities with factor level  $f_{iv}$  for factor  $f_i$  of vertex  $v$  and a binary function on factor levels  $s(f_{iv}, f_{iw}) \mapsto \{0, 1\}$  that denotes if two levels are similar or not, for example if both fall into a predefined interval  $I_x$ :

$$s(f_{iv}, f_{iw}) = \begin{cases} 1 & \text{if } f_{iv}, f_{iw} \in I_x \\ 0 & \text{otherwise} \end{cases} \quad (4.1)$$

where  $I_x = [x_a, x_b]$ ,  $x_a, x_b \in \mathbb{R}$ ,  $x_b \leq x_a$

We define a set of aggregation graphs  $H_1 \dots H_n$ , one for each factor  $f_i$ :

$$\begin{aligned} H_i &= (V_H, E_H) \\ \forall v \in V_G : \sum_{w \in V_H} s(f_{iv}, f_{iw}) &= 0 \rightarrow v \in V_H \\ \forall (v, w) \in E_G : v \in V_H \wedge w \in V_H &\rightarrow (v, w) \in E_H \end{aligned} \quad (4.2)$$

Each graph  $H$  summarizes all entities of  $G$  with a similar factor level into a single vertex, while preserving connections between the hierarchy levels of the experiment. For strings, that is, nominal factors, we define similarity as the perfect match of both levels (e.g., the same disease state), while quantitative variables can be summarized using intervals.

Figure 4.1 shows the aggregation of graph  $G$  based on factor  $f_1$  - two different treatments of mice - into graph  $H_1$ . At first, node  $v_1$  is added to the aggregation graph, since there is no node  $w$  in  $H_1$  with  $f_{1w} = f_{1v_1} = \text{treatment A}$ . Since  $f_{1v_2} = \text{treatment A}$ , the second node  $v_2$ , describing Mouse 2 is not added to the aggregation graph. Instead, the number of aggregated nodes for  $v_1$  is incremented. This process continues until all nodes in  $G$  have been visited. Notably, the more

complex levels of experimental factor  $f_2$  - the genotypes *wild type* and two different knockout variants - would lead to a more complex aggregation graph  $H_2$  consisting of nine nodes.

The principle used for experimental factor levels can be easily extended to aggregate the study graph based on different aspects of a study or on multiple aspects at once. For example, when including information about the organism or tissue in the boolean function, only entities from the same species or samples containing the same tissue types will be aggregated.

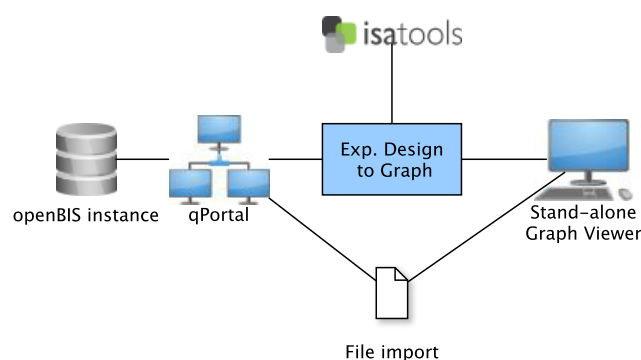
### 4.2.2 Validation on Studies

In addition to our own examples, we use a number of publicly available studies to test our approach. We selected ten studies to test the consistency of publically shared experimental designs with their related publications. We especially focused on the description of factor levels and the number of replicates and compared this information using the aggregation graph rendered from ISA-Tab. Studies were selected from the front page (latest studies) of MetaboLights, ignoring studies that were part of a larger publication including many more datasets as well as those studies belonging to publications that were not publically available. Data and metadata of the selected studies can be found in the MetaboLights database<sup>191</sup>. Information about the selected studies is available in Appendix B.

Additionally, we describe the design of one lipidomics study on the progression to islet autoimmunity and type 1 diabetes by Lamichhane et al.<sup>192</sup> more closely. It can be found using its identifier MTBLS620.

## 4.3 Design and Implementation

The schematic of the integration of our visualization into the portal can be seen in Figure 4.2.



**Figure 4.2:** Schematic diagram of our implementation: Existing experimental designs and attached metadata stored in openBIS can be visualized in qPortal using our Java-based experimental design libraries. Different formats can be imported and visualized, ISA-Tab being supported through isatools. An independent JavaFX implementation is available.

Both imported and existing experiments can be translated into the aggregation graph and displayed using Javascript libraries. For existing experiments, meta information about attached datasets is leveraged from the data store. When importing ISA-Tab investigations, the open-source framework isatools is used in the translation process, using source and sample identifiers of the ISA study format as well as all defined experimental factors as described in Chapter 3.4.4. The Javascript libraries dagre<sup>193</sup> and Data-Driven Documents ( $D^3$ )<sup>194</sup> are then used to compute graph coordinates and draw the selected graph. Interactions between Javascript and our portlets are implemented using remote procedure calls (RPC). A stand-alone version implemented in JavaFX can be used independently of the portal or openBIS.<sup>i</sup>

## 4.4 Results

### 4.4.1 qPortal Integration

Once the experimental design is parsed from a database or translated from a supported experimental design format, different experiment aggregation graphs can be drawn or redrawn in real-time by selecting different experimental factors. Factor levels and the type of analyte measured are the main discrimination criteria for aggregating similar nodes. However, the experiment graph can be drawn using different species or tissue types or any other property that is attached to samples or sources via metadata.

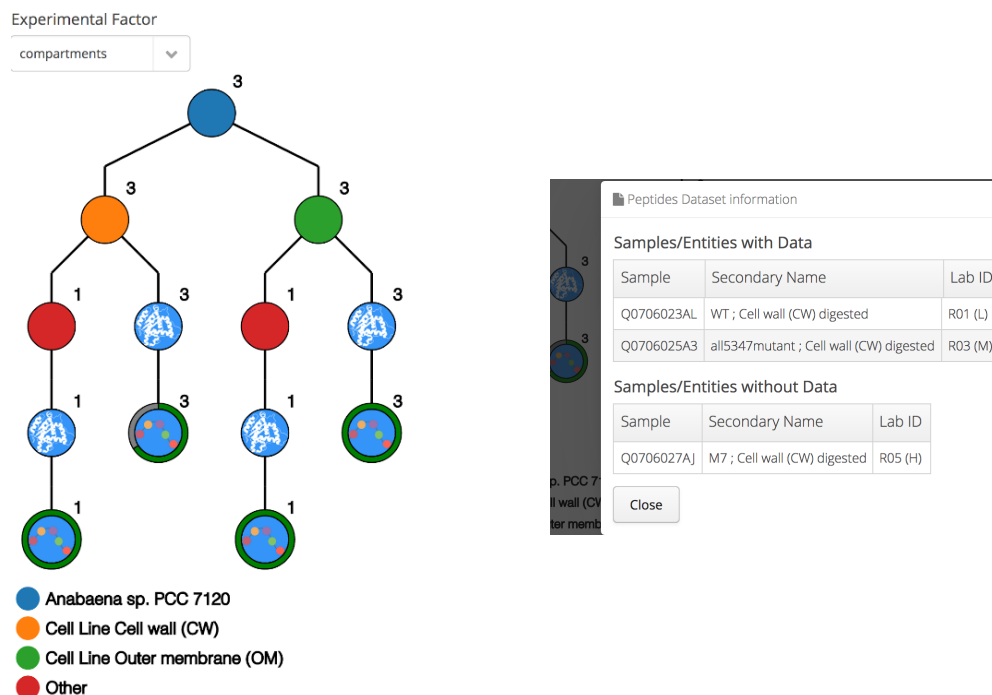
Further interaction is possible by clicking on nodes: dependent on the layer, users can display more information about sample sources, tissues or analytes. If a data store is attached, information about the status of the project, including which samples have already been measured and which data is still missing is displayed. This completeness of datasets is also visualized by a ring around the respective nodes, where the size of the colored arc denotes the fraction of samples for which data is available. Both of these features are shown in Figure 4.3.

Our implementation is not limited to single-omics experiments. Figure 4.4 shows a complex multi-omics study available as ISA-tab investigation, that has been imported. The experimental factor levels only differ between cell cultures, resulting in a graph that is rooted in the source species level.

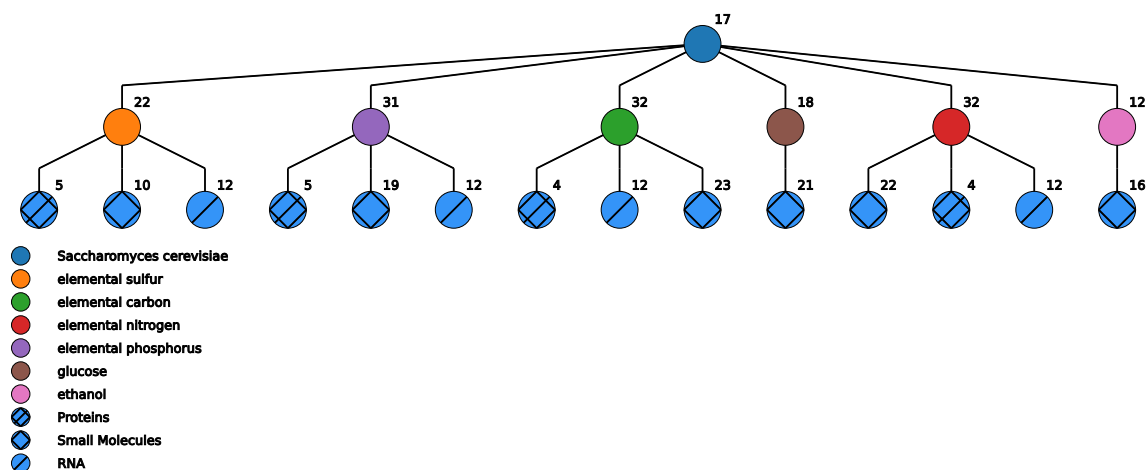
---

<sup>i</sup><https://github.com/qbicsoftware/experiment-graph-gui>

#### 4. Visual Exploration of Experimental Designs

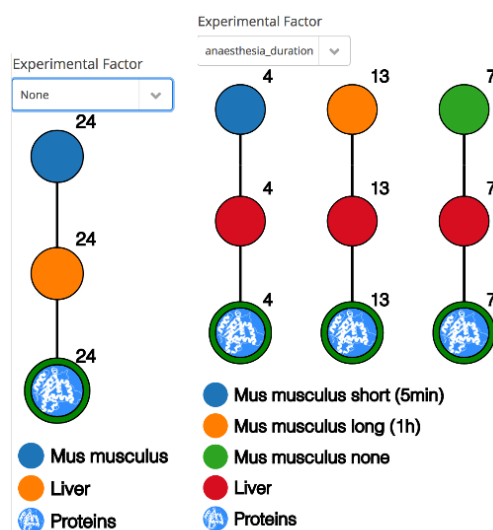


**Figure 4.3: Left:** experimental design graph representing an experiment on *Anabaena* cyanobacteria according to the experimental factor *compartments*. Two levels of analytes (proteins and peptides) illustrate proteomics-specific metadata collection of our model. The incomplete arc (grey part) on one of the nodes belonging to the *cell wall* level shows one dataset has not arrived in the database yet. **Right:** by selecting a graph node, a user has requested more information about the underlying samples. A popup table shows the respective identifiers and metadata of samples with and without attached data.



**Figure 4.4:** Aggregation graph of one study of an imported ISA-tab investigation. Yeast cultures are grown lacking different nutrients and proteome, transcriptome and metabolome are measured.

To compare our aggregation graph to the usual, complete sample hierarchy graph of a study, we demonstrate both visualizations on a proteomics experiment including 24 mice. All animals were anesthetized for different periods of time, liver tissue was extracted and proteins from those tissue samples were measured using mass spectrometry. Figure 3.4 shows a subset of the full sample graph without any experimental factors or levels displayed. In contrast, the aggregated experimental design graph of the same experiment seen in Figure 4.5 shows the condensed factor-independent overview, summarizing nodes by species, extracted tissues, and proteins.



**Figure 4.5:** Two experimental design graphs representing the same experiment on 24 mice. Left: no factor is selected, nodes are merged by species, tissue and analyte type. Right: the 3-level factor *anaesthesia duration* is selected, resulting in three subgraphs and the related legend explaining them.

Metadata like entity identifiers can be shown by clicking on nodes of the graph. The study can be explored further by selecting a factor of this experimental design from the drop-down menu. Selecting the factor *anaesthesia\_duration*, our algorithm splits the graph into three groups of mice and attaches descendant samples according to the three levels of this factor. Colors and legend of each aggregation graph are entirely dependent on the graph and inform users about species, tissues, analytes, and different factor values. Furthermore, the green outline of the protein nodes shows that data generation has been completed for all samples.

#### 4.4.2 Runtime

To examine the usability of our method for web-based applications, we compared the graph model creation time for studies of different sizes and experimental complexities. We also compare the different use cases of study import and displaying studies saved in our data

## 4. Visual Exploration of Experimental Designs

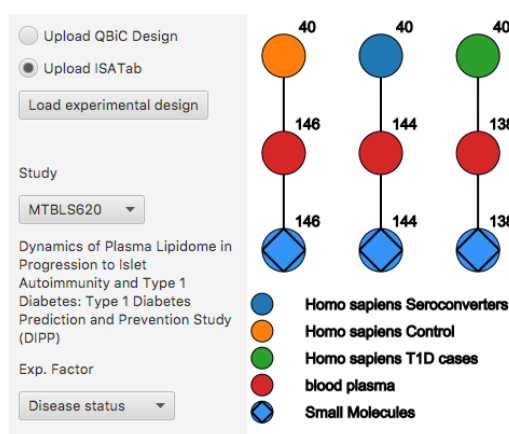
management system openBIS. Before studies using external formats can be registered, they have to be translated to our data model, leading to slightly longer runtimes. Despite this, Table 4.1 shows that run-time does not exceed 200 ms even for large studies.

**Table 4.1:** Aggregation graph creation time in milliseconds for studies of different numbers of entities, attached datasets, experimental complexity, and different formats. All tests were performed on a notebook using a 2.5 GHz Intel Core i5 processor.

No. of entities	No. of datasets	No. of factors	Runtime [ms]
qPortal studies			
424	224	4	30
716	575	4	56
1,076	5,361	1	51
2,202	475	3	85
Imported studies			
371	-	5	128
976	-	4	194
2,160	-	0	128

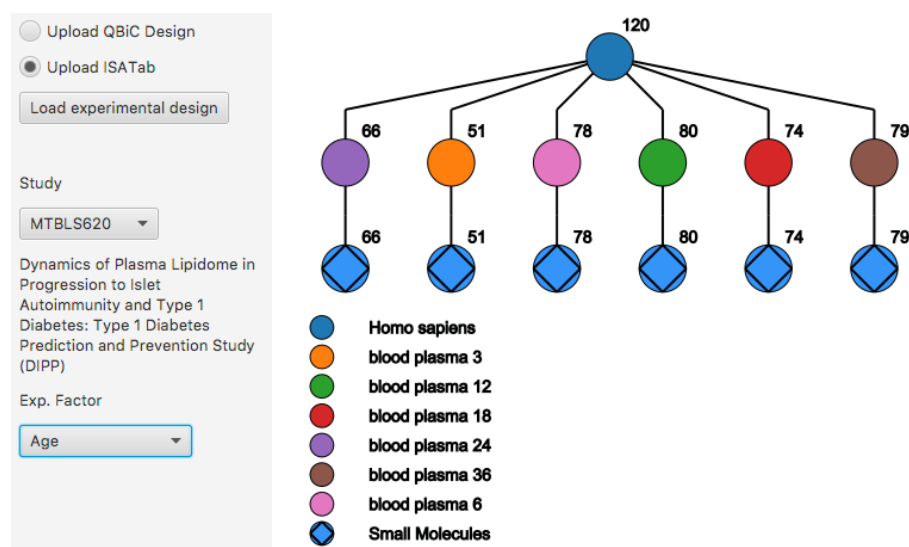
### 4.4.3 Validation

We evaluated our stand-alone implementation using a recent lipidomics study on the progression to islet autoimmunity and type 1 diabetes from the MetaboLights database. Our application shows a description of the imported ISA-study and lists every experimental factor that the authors have annotated in a drop-down menu. Selecting a study factor as seen in Figure 4.6 reveals more details about the study.



**Figure 4.6:** View of the stand-alone application after ISA-Tab import. The description of the selected study is displayed and users can select experimental factors (here: *disease status*) from a drop-down menu. Our aggregation graph shows extraction of blood plasma from 120 patients belonging to the three groups control, type 1 diabetes (T1D) and seropositivity (for islet cell autoantibodies). The metabolome of those samples was measured.

Selecting *disease status* shows that data generated from the blood plasma samples of 40 patients of a control group was compared to those of 40 type 1 diabetes (T1D) cases, as well as 40 cases of autoimmunity against islet cells, that had not yet progressed to diabetes. Selecting the *age* factor as seen in Figure 4.7 reveals that this is a time series study, where blood was taken at different subject ages. In this case, the authors failed to include units in their metadata, so it is only clear from their publication that the ages are measured in months.



**Figure 4.7:** View of the stand-alone application after selection of the experimental factor *age*. The levels of this factor show that blood plasma samples were taken at different ages of the same patients since the experimental levels are defined at the second level. The unit (months) is omitted since it was not annotated in the ISA-Tab study.

Using our visualization approach to test study design consistency, we found that for four studies (MTBLS618, MTBLS654, MTBLS669, MTBLS750), the described experimental factors and sample sizes found in the uploaded metadata fit the approaches described in the respective publications. For two additional studies, we found minor mistakes, such as describing biological replicates on the tissue level, while falsely specifying only one biological source (MTBLS619), or the ISA-Tab containing more biological sources and derived samples in addition to the ones specified in the publication (MTBLS780). Other studies are harder to reproduce, as the sample size in the shared metadata is smaller than described in the publication (MTBLS687), information about replicates is missing (MTBLS622), or neither factors nor replicates are described in ISA-Tab (MTBLS640, MTBLS674).

## 4.5 Discussion

We present tools to visualize large biomedical studies by their most important experimental aspects. Building on our graphical interface for the creation of factorial experimental designs and

our hierarchical data model, we create graphs summarizing complex hierarchies of experimental variables, allowing users to quickly familiarize themselves with the important aspects of a study. When used in a platform integrating experiment data and metadata, like qPortal, additional information about datasets can be leveraged, marking missing data or the status of a project. Our aggregation graph gives a concise and intuitive overview in cases where the representation of experiments was previously only possible using large tables.

The lack of statistical power and sound experimental design has led to the so-called reproducibility crisis. Extensive work has been done to standardize metadata annotation and storage, leading to public data repositories like PRIDE<sup>141</sup> or ENA<sup>2</sup> that aim to make proteomic and genomic studies findable and increase reproducibility and replicability in these fields. Other repositories, for example MetaboLights<sup>191</sup>, use multi-omics metadata standards in an attempt to unify the way scientists share their experimental data. Here, ISA-Tab provides a foundation to search, display and use the annotations found in its metadata model<sup>17</sup>. Such methods are clearly required due to the size of modern biomedical experiments and their metadata: a simple overview of a study often leads to huge tables or cluttered graphs. Some approaches are examples of successful, interactive uses of study design visualization, yet they address very specific questions. Bio-GraphIn<sup>189</sup> focuses on listing the replicates of each study group. By contrast, our approach provides an interactive visualization of a large number of experiments and is able to summarize replicates (with respect to one factor) into a single node to display a concise representation with which users can interact to control the displayed level of detail. Different approaches for graph aggregation<sup>195</sup> have been applied in multiple fields where information graphs are too large for human comprehension<sup>196,197</sup>. To our knowledge, these tools have not been used for the visualization of formats like ISA-Tab or graphs of biomedical experiments in general.

We have shown that our approach can display current studies including several hundred entities. Since ISA-Tab is not a minimum information standard, the amount of actual information beyond the sample hierarchy that can be drawn from its format depends on the annotations provided by researchers, as our examples show. We have taken the first steps towards a fully modular solution that will allow the integration of our tool in different contexts, fields of omics, and experiments. This can help enforce standards for various data models used by researchers.

Nonetheless, enforcing standards does not ensure the absence of annotation errors. Our comparison of publically available studies with their publications shows that the current approaches are not sufficient to guarantee the correct collection of meta information or even just the highly important study design. The majority of ISA studies we explored included at least minor errors in the definition of the experimental design. For some of these examples, we assess the metadata as sufficient to easily reproduce the experiments at hand. Multiple studies were, however, missing information about experimental factors or even the number of replicates, despite the respective publication stating the study design quite clearly. Many

discrepancies such as missing samples, study factors that are defined on the wrong level, or not at all, are much easier to spot using a visual representation of the study instead of large spreadsheets.

Experimental factors are one of the most important types of study annotation since they are at the core of the question scientists want to answer. However, our concept is not necessarily bound to the aggregation of different factor levels. Any property that can split subjects or samples into different groups, can be useful to find out more about a study. In large studies involving multiple groups, sharing information about the status of the project and data generation is often important. Provided this information is available, future work could include a time-component, displaying the history of a study.



## Chapter 5

# Interactive Sample Size Calculation for Differential Gene Expression Experiments

### 5.1 Introduction

In order to obtain reproducible results, scientific studies have to be built on a sound statistical foundation. An important part of this process includes using an adequate number of replicates as described in Section 2.2.2. For contemporary science, estimating the needed sample size to detect an effect is even more important, as high-throughput biomedical experiments can be costly. Time and money are lost if wrong conclusions or none at all can be drawn from data that is too noisy. On the other hand, an experiment with more subjects than needed might be unethical, as well as wasteful.

For these reasons, multiple approaches have been developed to estimate the statistical power of experiments generating biological high-throughput data. For quantitative proteomics, earlier methods focused on the sharing of methodology to plan powerful experiments<sup>198–200</sup> or the use of generalized software tools, which were not specifically developed for the field<sup>198,201,202</sup>. With these approaches, responsibility generally fell to the user to provide the software with a measure of variance from older experiments, in order to compute the power of a newly planned experiment. More recently, tools specific to mass spectrometry have been developed. The R package MSstats enables sample size estimations for MS-based proteomics data<sup>203</sup>. MSstats has since been integrated into the OpenMS framework for MS-based analysis and application development<sup>204</sup>. For the field of differential gene expression, there have been many approaches tailored to cDNA microarrays<sup>205–207</sup> as well as RNA sequencing<sup>208,209</sup>.

Nevertheless, especially for the budding field of RNA-Seq, there is yet no accepted standard on how deep an organism needs to be sequenced or on how to estimate the in-group variance of gene expression in order to predict these statistics. While the methods based on negative binomial models, as used by tools like edgeR and DESeq 2, show promising results, they are often only implemented as scripts to be used on the command line or as packages for statistical tools or languages. This restricts their reach to scientists familiar with R or other programming languages with a heavy focus on statistics.

For these reasons, there have been various efforts to provide researchers with web interfaces to design differential gene expression experiments<sup>210,211</sup>. The tool Scotty uses a Poisson model to enable users to plan their RNA-Seq experiments subject to different criteria like cost and power<sup>212</sup>. RnaSeqSampleSize leverages the web technology Shiny to provide a web interface to the statistical computations performed in the background<sup>213</sup>. Its users can upload pilot data to compute the necessary parameters or give their own estimations. The sample size and power estimations are based on the negative binomial model used by many modern tools for differential expression analysis.

Unfortunately, some of these tools are no longer available to researchers. Additionally, existing solutions are more or less decoupled from the experimental process consisting of metadata and data collection, and data analysis. Users have to choose one model and make do with the information they have. There is no tool that leverages existing metadata to help users arrive at a robust experimental design, despite information about the experiment's organism and type of study performed being some of the issues at the core of this process. Here, we present an integrative approach based on the R packages OCplus for DNA microarray power estimation<sup>128</sup> and RnaSeqSampleSize for RNA-Seq power estimation, as well as our own web portal. The process leverages data sets and experimental design meta information stored in our system. Users are provided with a straight-forward web interface where they can select options to estimate different parameters based on literature or publicly available datasets in order to obtain visualizations for sample size or power of their experiments.

## 5.2 Material and Methods

In order to predict the required sample size and various statistical power measures for Microarray experiments, we use the R package OCplus based on the work by Pawitan et al.<sup>128</sup>. Given sample size and log fold change, the mixture model approach allows the prediction of false discovery rate as well as false negative rate based on a percentage of genes that are declared as differentially expressed. We extend this method by computing the relevant statistical measures for different parameters for sample size and log fold change, creating different power matrices that can be displayed to the user.

### 5.2.1 RNA-Seq Variance Model

For RNA-Seq, we base our estimations on the negative binomial distribution as it is widely used by tools like edgeR, DESeq 2, and consequently RnaSeqSampleSize. As described in Chapter 2, the relative standard deviation — which is also known as the coefficient of variation  $CV$  — of a gene expression count is solely dependent on the inverse of the mean of counts  $\frac{1}{\mu}$  of the respective gene and on the dispersion  $\alpha$ . As read counts are dependent on the sequencing

depth,  $\alpha$  is also known as the biological coefficient of variation<sup>129</sup>. The deeper samples are sequenced, the larger the mean and the lower the effect of technical variation on the coefficient of variation.

Hence, both the average read count and the dispersion must be estimated in order to predict sample size and power of an experiment. The `RnaSeqSampleSize` package includes publicly available datasets from the cancer genome atlas (TCGA)<sup>214</sup> and allows estimation for these and other raw datasets provided by users. Additionally, we include estimations of average read count and dispersion from literature<sup>215,216</sup> and general estimations of square-root-dispersion for human data, for that of genetically similar model organisms, and for technical replicates as they are found in the edgeR user guide<sup>217</sup>.

### 5.2.2 Parameter Optimization and Sample Size Calculation

`RnaSeqSampleSize` provides functionality to create and visualize power matrices in R, allowing the optimization of two parameters in respect to resulting sample size or statistical power. Here, we focus on two use cases. First, we estimate power for a known experimental design and study factor. For unbalanced study designs, the smaller number of samples is used. Based on the known or estimated ratio of DE genes, dispersion and average read count of diagnostic genes, a power matrix is created using several values for *FDR* and the *minimum fold change* between two groups of the study factor. Our second use case for RNA-Seq is sample size estimation based on the ratio of DE genes, dispersion, FDR level and an estimate for average read count. Here, the parameters *statistical power* and *minimum fold change* are varied in the optimization process to create a sample size matrix.

For DNA microarray power analysis and sample size estimation, the `OCPlus` R package can plot power or sample size against FDR/FNR. We adapt these methods in order to create power and sample size matrices for our two use cases. This first approach is based on the minimum statistical power a researcher wants to reach for a predefined experimental design containing a study factor including samples of two groups. For this selected sensitivity and known sample size, we vary *log fold change* and *the ratio of non-DE genes* in order to predict the FDR for this statistical power level. For sample size estimation, we optimize the detectable *log fold change* and *sample size* in order to show FDR/FNR based on declaring a certain ratio of genes as differentially expressed.

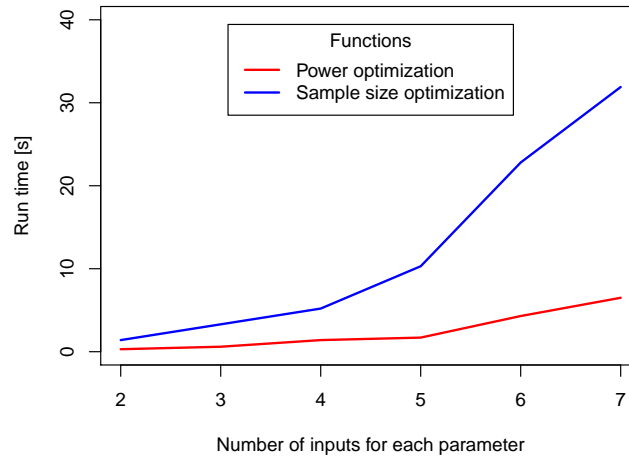
## 5.3 Implementation

Our Vaadin portlet uses the Java library `RVaadin`<sup>218</sup> to provide interactive R functionality for DNA microarray statistics. `RVaadin` allows the connection to a binary R server (`Rserve`) in order to communicate function calls and results. We use R scripts to enhance the package

OCPlus by running multiple estimations for each use case and creating a power or sample size matrix, displaying the respective measure as a heatmap. For RNA-Seq power estimation, we additionally provide the option to integrate pilot count data from our system to estimate gene dispersion or average read counts. A connection to our data management system openBIS allows for the integration of existing experimental design information and storing of results.

### 5.3.1 Backend

Unlike the optimizations using OCPlus, the runtime for sample size predictions on RNA-Seq data is prohibitive for an interactive approach in most cases, as seen in Figure 5.1 and even larger for parameter-estimation on real data.



**Figure 5.1:** Runtime of power and sample size estimation functions using `RnaSeqSampleSize` without read count data for different input parameter sizes. Two parameters are optimized in each case, resulting in quadratic growth of the power/sample size matrix.

Therefore, we created a Singularity<sup>168</sup> container to perform these computations on a virtual machine. This enables us to provide multi-parameter optimization using the package `RnaSeqSampleSize` to create power and sample size estimation matrices based on TCGA datasets or other RNA sequencing runs.

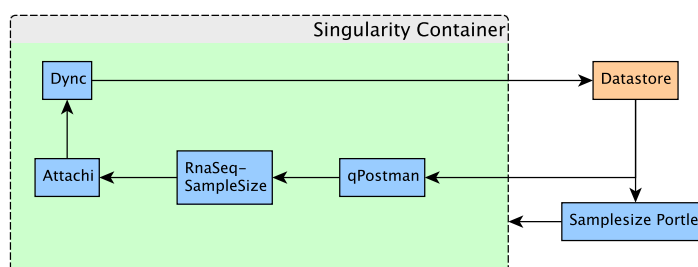
Selecting the respective options via the portlet registers meta information about the run in openBIS and executes the container via a Secure Shell (SSH) call to the virtual machine. If read count data is used, the respective dataset is downloaded to the VM using qPostman<sup>i</sup>. Power matrices are then created using the `RnaSeqSampleSize` package. A script (attachi<sup>ii</sup>)

---

<sup>i</sup><https://github.com/qbicsoftware/postman-cli>

<sup>ii</sup><https://github.com/qbicsoftware/attachi-cli>

creates the necessary information needed to register the data in openBIS. Results are securely transferred via Dync<sup>iii</sup>, triggering the data registration process. The selected parameters and the resulting matrix image are attached to the respective project selected by the user when starting the computations and can be displayed via multiple of our portlets. Figure 5.2 shows the interplay between the different tools contained in our Singularity container and the data storage.



**Figure 5.2:** Schematic view of the Sample Size portlet backend. Power or sample size estimation for RNA-Seq is performed using a singularity container on a virtual machine. If read count data from openBIS is used, the tool qPostman stages the respective datasets. Results are written back to openBIS using the tools attachi and Dync. Users can then find the generated information in their project space.

## 5.4 Results

Our Singularity container and its source code is available on GitHub<sup>iv</sup>. The graphical user interface is available on qPortal<sup>v</sup>.

### 5.4.1 Graphical User Interface for Power Analysis of Experimental Designs

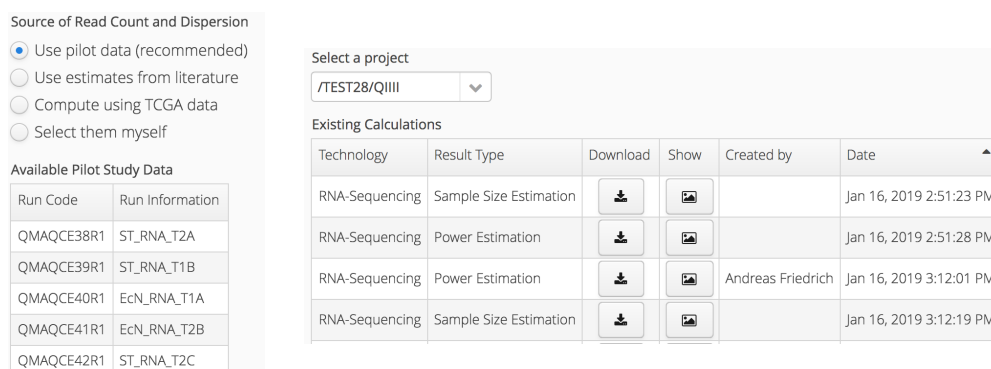
In order to make use of data and metadata registered via qPortal, we present users with a project-centric user interface. While sample size estimation for DNA microarrays can be performed outside of the project context, the extended functionality of the portlet is only available once a project has been selected. Figure 5.3 shows the presentation of existing read count and result data related to a selected project.

<sup>iii</sup><https://github.com/qbicsoftware/dync-cli>

<sup>iv</sup><https://github.com/qbicsoftware/rnaseq-power-container>

<sup>v</sup><https://portal.qbic.uni-tuebingen.de/portal/web/qbic/samplesize>

## 5. Interactive Sample Size Calculation for Differential Gene Expression Experiments



**Source of Read Count and Dispersion**

☒ Use pilot data (recommended)  
☐ Use estimates from literature  
☐ Compute using TCGA data  
☐ Select them myself

**Available Pilot Study Data**

Run Code	Run Information
QMAQCE38R1	ST_RNA_T2A
QMAQCE39R1	ST_RNA_T1B
QMAQCE40R1	EcN_RNA_T1A
QMAQCE41R1	EcN_RNA_T2B
QMAQCE42R1	ST_RNA_T2C

**Select a project**

/TEST28/QIIII

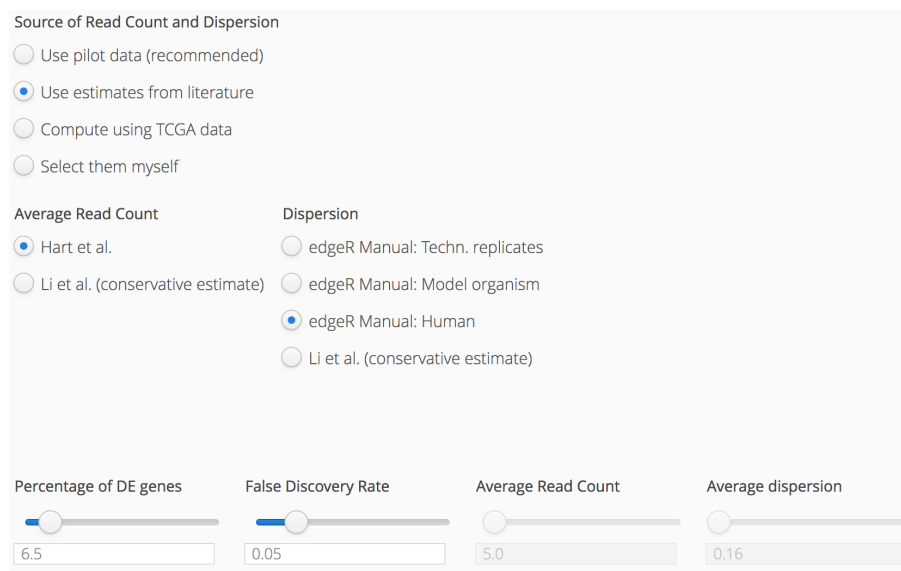
**Existing Calculations**

Technology	Result Type	Download	Show	Created by	Date
RNA-Sequencing	Sample Size Estimation				Jan 16, 2019 2:51:23 PM
RNA-Sequencing	Power Estimation				Jan 16, 2019 2:51:28 PM
RNA-Sequencing	Power Estimation			Andreas Friedrich	Jan 16, 2019 3:12:01 PM
RNA-Sequencing	Sample Size Estimation				Jan 16, 2019 3:12:19 PM

**Figure 5.3:** Presentation of project-specific data in the portlet. **Left:** Gene dispersion and average read count can be estimated from RNA-Seq data found in openBIS. **Right:** Previous sample size or power estimation runs for a project can be displayed or downloaded.

If datasets containing raw count data from RNA-Seq runs are available, for example as part of a pilot project, they are displayed when selecting the respective prediction option. Additionally, previous sample size or power estimation runs are displayed in a table. This data can be downloaded via the click of a button or displayed directly in the browser alongside the parameters of the prediction.

Figure 5.4 shows some of the available parameter options for RNA-Seq sample size estimation.



**Source of Read Count and Dispersion**

☐ Use pilot data (recommended)  
☒ Use estimates from literature  
☐ Compute using TCGA data  
☐ Select them myself

**Average Read Count**

☒ Hart et al.  
☐ Li et al. (conservative estimate)

**Dispersion**

☐ edgeR Manual: Techn. replicates  
☐ edgeR Manual: Model organism  
☒ edgeR Manual: Human  
☐ Li et al. (conservative estimate)

**Percentage of DE genes**

6.5

**False Discovery Rate**

0.05

**Average Read Count**

5.0

**Average dispersion**

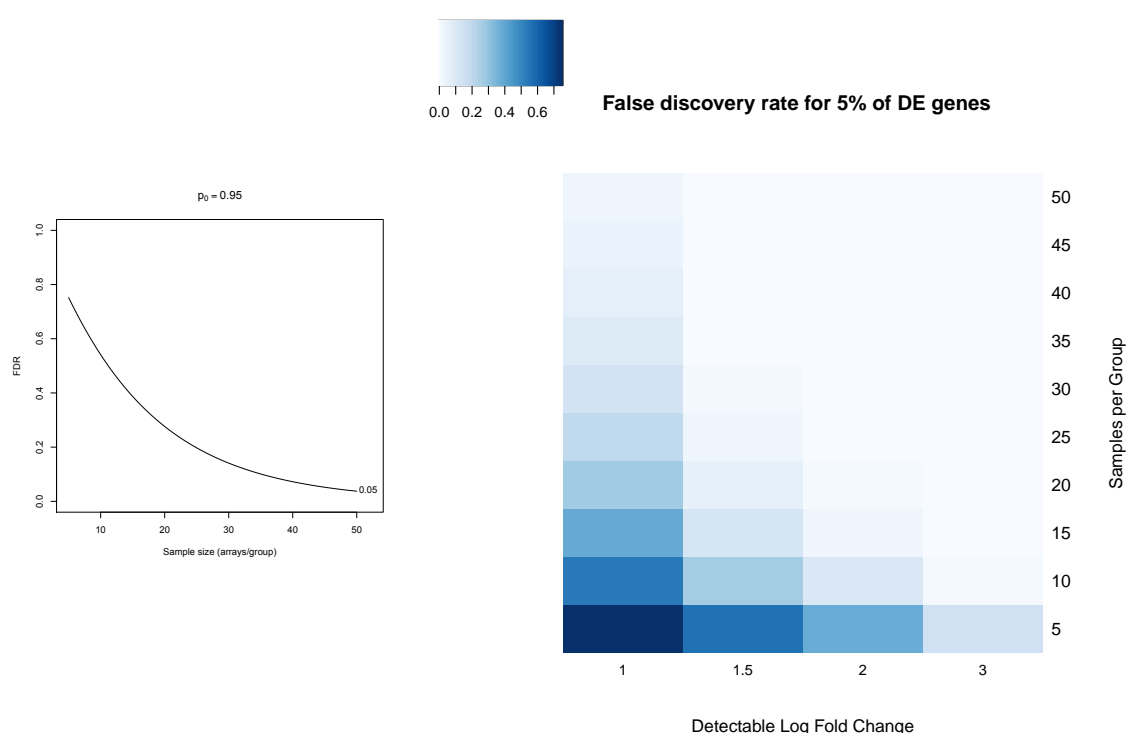
0.16

**Figure 5.4:** Parameter selection for RNA-Seq sample size prediction using our portlet. All parameters can be selected via sliders or input fields. For average read count and dispersion, several other options are available. Both can be estimated from provided pilot data, TCGA cancer datasets or based on estimates from literature. In the case shown here, an average read count of 5 has been selected based on Hart et al.<sup>216</sup> and dispersion was set to 0.16 based on the square-root-dispersion for human data proposed by Chen et al.<sup>217</sup>.

All parameters can be chosen via a slider or the respective input field. Additionally, gene dispersion and the average read count of genes can be selected based on literature, one of the included TCGA datasets, or data available in openBIS.

### 5.4.2 Validation

We validated our implementation by comparing our results to the output of the original R packages. Figure 5.5 shows a comparison of sample size estimation when 95% of genes are assumed to be non-differentially expressed and the top 5% of genes are reported as differentially expressed.



**Figure 5.5:** Comparison of FDR estimation for selection of the top 5% DE genes, assuming 95% are non-DE. **Left:** OCPlus plot of FDR as a function of sample size for log FC = 1. **Right:** Our portlet creates a matrix displaying the FDR levels as a function of both sample size and log FC. The left column corresponds to the OCPlus plot.

OCPlus plots FDR against sample size for a specific log fold change detection level. This corresponds to a single column of our FDR matrix. This is shown in more detail in Table 5.1. We can see, that five arrays per group produce a high false discovery rate of 75% for genes with a log fold change of 1 and 38% for genes with a log fold change of 2. 45 arrays per group would be needed in order to decrease the FDR to 5%.

## 5. Interactive Sample Size Calculation for Differential Gene Expression Experiments

**Table 5.1:** Complete FDR table for parameter optimization using four different log fold changes and various arrays per group as considered parameters.

Log Fold Change	1	2	3	4
Arrays/group	FDR			
5	0.75	0.38	0.15	0.06
10	0.54	0.11	0.01	$1.52 \times 10^{-3}$
15	0.39	0.03	$1.28 \times 10^{-3}$	$4.34 \times 10^{-5}$
20	0.28	$9.75 \times 10^{-3}$	$1.26 \times 10^{-4}$	$1.30 \times 10^{-6}$
25	0.20	$2.97 \times 10^{-3}$	$1.27 \times 10^{-5}$	$4.01 \times 10^{-8}$
30	0.14	$9.16 \times 10^{-4}$	$1.29 \times 10^{-6}$	$1.26 \times 10^{-9}$
35	0.10	$2.84 \times 10^{-4}$	$1.34 \times 10^{-7}$	$4.00 \times 10^{-11}$
40	0.07	$8.88 \times 10^{-5}$	$1.39 \times 10^{-8}$	$1.28 \times 10^{-12}$
45	0.05	$2.79 \times 10^{-5}$	$1.45 \times 10^{-9}$	$4.22 \times 10^{-14}$
50	0.04	$8.78 \times 10^{-6}$	$1.53 \times 10^{-10}$	0.00

For validation of our RNA-Seq implementation, we successfully compared the most complex use case, power estimation based on sequencing data, to the RnaSeqSampleSize implementation. Table 5.2 shows the resulting statistical power for different combinations of fold changes and false discovery rate for a sample size of  $n = 50$  and an estimated 5% of differentially expressed genes.

**Table 5.2:** Complete power table showing parameter optimization for the breast invasive carcinoma dataset (BRCA) of TCGA using different fold changes and different FDR thresholds. Sample size was  $n = 50$ ,  $p_0 = 0.95$  and 100 genes were used in the analysis.

Fold Change	1.1	1.5	2	3
FDR	Power			
0.01	$3.50 \times 10^{-5}$	0.26	0.70	0.87
0.03	$1.21 \times 10^{-4}$	0.33	0.73	0.88
0.05	$2.14 \times 10^{-4}$	0.37	0.74	0.89
0.10	$6.92 \times 10^{-4}$	0.42	0.76	0.90
0.20	$2.37 \times 10^{-3}$	0.48	0.78	0.92

100 genes from the TCGA dataset for breast invasive carcinoma (BRCA) were used in the estimation. For this dataset and sample size, differentially expressed genes with a low fold change of 1.1 cannot reliably be predicted even for relatively high FDR thresholds (power of 0.002 for 20% FDR). A modest percentage (26-37%) of DE genes with fold changes of at least 1.5 can be found even for strict FDR cutoffs between 0.01 and 0.05 and DE genes with fold changes of 2 or higher lead to at least 70% sensitivity.

## 5.5 Discussion

Sample size estimation and power analysis are an important part of modern omics projects. Besides general guidelines<sup>198–200</sup> or software for statistical estimations<sup>201,202</sup>, many omics-specific approaches and tools have been developed in recent years<sup>203,205–212</sup>. However, the large majority either lack an intuitive user interface or are not integrated into existing software systems, where additional experimental data or metadata would be of great help for the analysis process. Here, we present our integration of two R packages for DNA microarray<sup>128</sup> and RNA-Seq<sup>213</sup> statistical power estimation into our web portal. This approach allows us to leverage data sets and meta information that is stored in our system. Users are provided with a straight-forward web interface with which they can select options to estimate different parameters based on literature or publicly available datasets in order to obtain visualizations for sample size or power of their experiments. For microarray experiments, we augment the existing OCplus package in order to provide power and sample size matrices and display them directly in our portal. For RnaSeqSampleSize we provide the ability to automatically process pilot read count data from our system on a virtual machine in order to predict sample size or statistical power. For both approaches, our data model allows users to explore the power of existing experimental designs. Study factors of interest can be selected and the number of biological replicates per factor level is automatically used for predictions. Additionally, the input parameters of each power analysis are stored, displayed to users of the portlet and can be downloaded alongside the prediction results. OCplus was chosen as it presents an interesting alternative to the classical statistical testing paradigm of focussing on the type I error rate. Depending on the aim of a study, higher sensitivity and therefore less false negatives might be more promising than keeping the false discovery rate low. Researchers using our portlet are empowered to select the preferred minimum sensitivity for their planned study and can then determine the number of needed replicates to stay below a certain FDR and FNR cutoff value. Similarly, some studies only target genes with very high fold changes (e.g., when searching for potential biomarkers), and might not require the power to detect most of the differentially expressed transcripts. Through visualization of our power matrix, the influence of different possible log fold changes on sensitivity and FDR can be easily predicted without performing multiple runs of the basic version of OCplus. Alternatively, the relationship between power and FDR for existing experimental designs can be discovered by selecting the two factor levels that are to be compared. Here, the impact of different DE ratios is reported and can help scientists estimate the power of their respective studies more confidently. In all of these cases, the information is presented interactively, allowing for fast exploration of the statistical properties of different designs. One disadvantage of OCplus is the missing modeling of biological variation. Pawitan et al.<sup>128</sup> base their model on the ratio of truly differentially expressed genes, which is an implicit measure of this variation. However, this parameter can

be hard to predict outside of studies using model organisms that typically show high genetic homogeneity. OCplus does, however, allow the estimation of this parameter from pilot data and we are working on including this feature.

For RNA-Seq power analysis, we selected the more recently published package `RnaSeqSampleSize` since it is based on the same statistical foundation as commonly used analysis tools for RNA sequencing. The question of how deep transcriptomes need to be sequenced in order to maximize statistical power for differential expression experiments has previously been discussed<sup>37,38</sup>. Liu et al.<sup>39</sup> have shown that increasing sequencing depth beyond 10 million reads for human samples leads to diminishing returns and money can often be saved by instead using more biological replicates. However, the correct strategy is also dependent on specific transcripts of interest and their abundance. While 10 million reads with sufficient replicates lead to a powerful experiment overall, low-abundant transcripts might very well be missed in the analysis. `RnaSeqSampleSize` and our interface enable users to specify their estimated average read count in order to receive estimations that are more closely tailored to the experiment at hand.

In contrast to OCplus, `RnaSeqSampleSize` also takes biological variation in the form of dispersion estimates into account by building on the negative binomial model for sequenced reads. The most accurate way to predict dispersion is to generate pilot data using the same biological system, conditions, and protocols as in the planned experiment. Our interface provides an easy way to select the correct datasets for this. Unfortunately, this approach is not yet followed frequently. To enable predictions without pilot data, `RnaSeqSampleSize` allows estimating this information from a number of TCGA datasets based on different human cancer samples. Additionally, we provide estimates for different organisms from studies and the edgeR guide. The annotated data available on our portal present an opportunity for future research into tissue- and organism-dependent biological variation. Since technical variation in RNA-Seq is very low and typically controlled for by sequencing depth<sup>129,219</sup>, such data could provide useful benchmarks for future studies without pilot data.

Power estimation of experiments remains a crucial topic in many fields of high-throughput biomedical research. New methods to generate data faster and more precisely will require the adaption or the creation of new models and tools to plan statistically sound experiments. Our data model and the modular structure of our portlet and backend allows the integration of statistical power analysis tools for additional omics technologies. Our implementation uses `RVaadin` for fast, interactive statistical estimations. Power estimation tools for complex high-throughput technologies often simulate or even partly perform the respective analysis, potentially leading to long run times. This complicates existing online user interfaces and can be detrimental to the user experience. Our approach solves this problem in two ways. First, the complete computation process is handled on a separate server, enabling the user to perform other tasks in our portal. Secondly, the server's connection to our data management

system allows users to end the session and come back later to find their completed results. Furthermore, the approach makes use of versioned Singularity containers and collects metadata of the used parameters. Both are freely available to the user, complying with the FAIR data principles.



## Chapter 6

# Using Experimental Design for Data Processing and Visualization

### 6.1 Introduction

Aspirin is the first and one of the most well-known nonsteroidal anti-inflammatory drugs (NSAIDs). Apart from its analgesic and antipyretic effects, it has become an important drug for the treatment and prevention of cardiovascular diseases as it inhibits platelet aggregation<sup>220,221</sup>. Despite these benefits, its application is complicated by human genetic variation. Thus, effective doses vary between patients and a significant part of the population shows aspirin resistance, greatly reducing the efficacy of the drug<sup>222,223</sup>. This has posed significant problems with aspirin dosage, especially when considering that the downside of reduced blood coagulation can be increased susceptibility to and greater severity of bleeding<sup>221,224</sup>. These effects have led researchers to develop rapid tests for aspirin susceptibility. The different human phenotypes also provide an opportunity to study the genetic causes of aspirin resistance and the pathways involved in platelet function, the so-called aspirin response signature (ARS). As with any research topic, the ability to reproduce previous work is an important tool to falsify, confirm or re-use results of aspirin-related studies. As we have argued in the previous chapters, one of the most important requirements for reproducibility of results is the availability of metadata, especially of the experimental design a study is based on. It is equally important to provide raw data to other researchers in order to reproduce the steps of a study as closely as possible since each of them represents a unique transformation of data, that can change the end results. Most importantly, raw data enables researchers to perform quality control. For DNA microarray studies, a large variety of visualization methods exist that can help to assess the quality of normalization, the clustering of experimental levels, or other effects that may show errors in array preparation and may have detrimental effects on the data analysis.

Here, we present the application of the work described in the previous chapters, to examine the reproducibility of existing research on the aspirin response signature. We imported a publicly available DNA microarray dataset and its metadata into our portal, leveraging our model for experimental designs described in Section 3.3. We performed interactive statistical power analysis using information taken from the experimental design and discuss the implications of

our approach for further analysis. Using the integration of our metadata model and the stored data, we developed a normalization and quality control workflow for our portal that automates important steps in the analysis of microarray data, while at the same time allowing users to select the experimental factor that should be used to display quality measures. Following the quality control, we use various approaches to detect differential expression in order to compare our results with the original study.

### 6.2 Materials and Methods

We make use of a DNA microarray dataset of 26 human patients<sup>225</sup> showing varying responses to aspirin according to their Aspirin Reaction Units (ARU) as measured with the VerifyNow test<sup>226</sup>. Subjects were divided into aspirin-resistant (AR, > 550 ARU), high normal (HN, ARU 500-549), and aspirin-sensitive (AS, ARU < 500) groups. Samples were taken from whole blood after 7-10 days of continuous aspirin administration (~81 mg/day). RNA expression was measured on the Affymetrix U133 Plus 2.0 microarray platform. Study metadata in the ISA-Tab format was imported from the Personalized NSAID Therapeutics Consortium (PENTACON, [www.pentacohq.org](http://www.pentacohq.org)) study E-GEOD-38511. Associated Affymetrix data was downloaded from the Gene Expression Omnibus dataset GSE38511.

#### 6.2.1 Quality Control Workflow

We implemented an R-based quality control workflow for Affymetrix arrays that makes use of captured experimental design information. Our workflow computes and visualizes various quality measures like comparative intensity box plots before and after normalization based on the experimental factor selected by a user. Group-wise MA plots show bias between arrays of different factor levels<sup>106</sup>. Visualization of the principal components can show clustering of samples by factor levels or indicate outliers. Similarly, hierarchical clustering of expression values is visualized using the package dendextend<sup>227</sup> in order to allow exploration of the data.

Use of the grid and cloud user support environment (gUSE<sup>228</sup>) as described by Mohr et al.<sup>186</sup> enables parameter selection, data staging, and execution. Results of our workflow are transformed to HTML and can be visualized in our Liferay portal or downloaded with additionally included QC data.

#### 6.2.2 Differential Expression Analysis

Arrays were normalized using the robust multichip average (RMA) method. Differential expression analysis was performed using two-sided Student t-tests between the samples of aspirin-resistant (AR) and all other patients (NR), consisting of aspirin-sensitive (AS) and high normal (HN), as in the original study by Fallahi et al.<sup>225</sup>. Additionally, the limma package<sup>229</sup> was

used to compare results. For both methods, genes were reported as significantly differentially expressed based on the criteria laid out by the original study. The authors used a fold-change cut-off of at least 1.5 combined with p-values below the threshold of  $\alpha = 0.001$ . In addition to this comparison with the original study, we performed multiple testing correction using the Benjamini-Hochberg<sup>124</sup> method using a threshold of  $\alpha = 0.05$ .

We note, that when the authors state a fold change of 1.5, they refer to the quotient of the larger group mean divided by the smaller group mean, also including down-regulated transcripts showing the respective fold change not exceeding  $\sim 0.67$ . In the rest of this work, we follow the same terminology when talking about the fold change cut-off value. For our results, we report log fold changes and denote the direction of differential expression by a sign.

For a direct comparison between the groups of AR and AS patients, we used the same tools, but a different strategy. We estimated the proportion of non-DE genes by fitting the mixture t-distribution to the vector of observed t-statistics between the two groups. This was performed using the *tMixture* statistic of the OCplus package. Based on the result we selected the top 3.5% genes with the smallest p-values according to our power estimation based on the model by Pawitan et al.<sup>128</sup>.

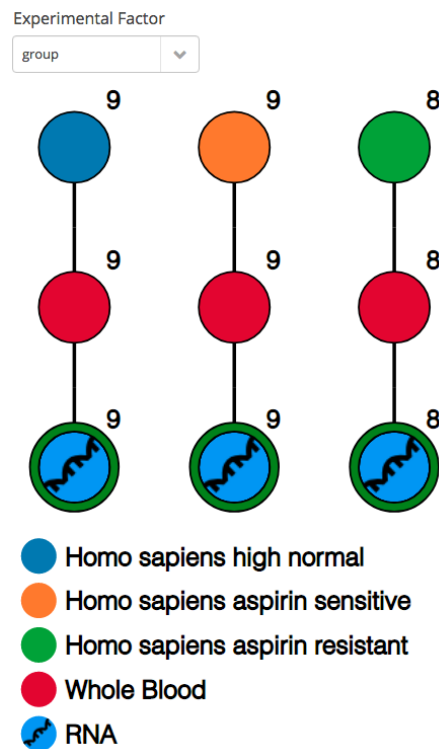
All computations relating to differential expression analysis were performed using R version 3.5.1 on the x86\_64-apple-darwin15.6.0 platform unless stated otherwise.

Gene set enrichment and functional analysis were performed with the web-platform DAVID version 6.8<sup>230</sup> using Affymetrix probe set identifiers. We added the GAD disease database to the standard selection but used default settings in every other category.

## 6.3 Results

### 6.3.1 Study Import

Study metadata was imported via the ISA-Tab format, creating unique openBIS sample identifiers in the process. Identifiers were mapped to the GEO identifiers captured from the ISA-Tab assay file and attached to the file names of the raw data. Files were then uploaded to the data store and automatically attached to the respective samples. Successful data registration for all three factor levels can be seen in Figure 6.1.

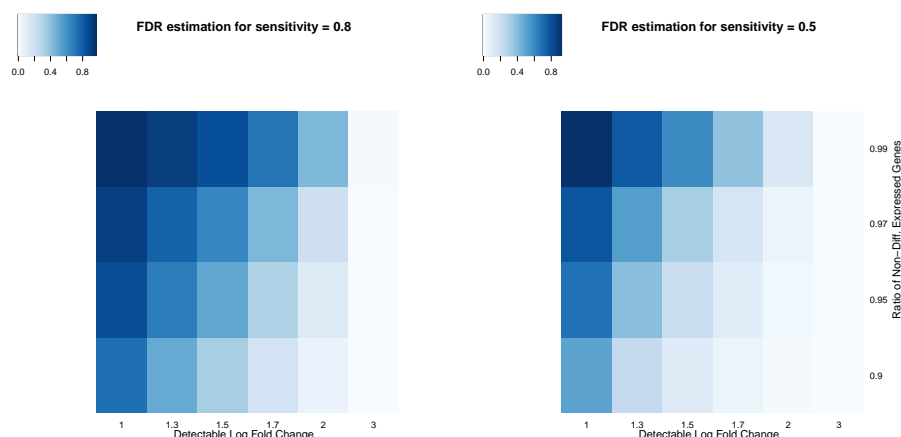


**Figure 6.1:** Graph of the aspirin study using the factor *group* after data has been registered. The experimental factor *group* was automatically imported from ISA-Tab, enabling a grouping by different levels and an overview of the related sample sizes. Completed green arcs denote that all raw data has been registered.

### 6.3.2 Sample Size Estimation

Based on the registered experimental design of the study, we performed an estimation of FDR for the comparison between samples from AR patients ( $n = 8$ ) vs. AS patients ( $n = 9$ ). Figure 6.2 shows that the false discovery rate is consistently large ( $\geq 0.50$ ) for differentially expressed genes with only low effect sizes (log fold change of 1). For the predicted ratio of non-DE transcripts, FDR is 0.89 for a sensitivity threshold of 0.8. Even if we accept a sensitivity of only 0.5, the FDR is only marginally lower (0.75).

For transcripts with a log fold change of 2 and a sensitivity of 0.8, the respective FDR prediction is 0.18. It is 0.04 if we accept that we will miss half of the DE transcripts (sensitivity of 0.5).



**Figure 6.2:** Matrix of FDR estimations for different log fold changes and ratios of DE genes for aspirin-resistant ( $n_1 = 8$ ) and aspirin-sensitive patients ( $n_2 = 9$ ). **Left:** Estimations for a sensitivity of 0.8 **Right:** Estimations for a sensitivity of 0.5.

### 6.3.3 Quality Control

We applied our workflow to all 26 datasets of the the aspirin study using the workflow interface of qPortal<sup>186</sup> as seen in Figure 6.3. The factor patient group was selected as a parameter.

► Submission: Microarray QC

Select input file(s)

InputFiles.1.input

<input checked="" type="checkbox"/>	File Name	File Type
<input checked="" type="checkbox"/>	QHASP050AQ_GSM943896_AR081.CEL	Q_MA_RAW_DATA
<input checked="" type="checkbox"/>	QHASP051A0_GSM943878_AR075.CEL	Q_MA_RAW_DATA
<input checked="" type="checkbox"/>	QHASP052A8_GSM943875_AR089.CEL	Q_MA_RAW_DATA
<input checked="" type="checkbox"/>	QHASP027AR_GSM943876_AR109.CEL	Q_MA_RAW_DATA
<input checked="" type="checkbox"/>	QHASP028A1_GSM943882_AR027.CEL	Q_MA_RAW_DATA
<input checked="" type="checkbox"/>	QHASP029A9_GSM943894_AR060.CEL	Q_MA_RAW_DATA
<input checked="" type="checkbox"/>	QHASP030AC_GSM943883_AR015.CEL	Q_MA_RAW_DATA
<input checked="" type="checkbox"/>	QHASP031AK_GSM943895_AR067.CEL	Q_MA_RAW_DATA
<input checked="" type="checkbox"/>	QHASP032AF_GSM943874_AR073.CEL	Q_MA_RAW_DATA

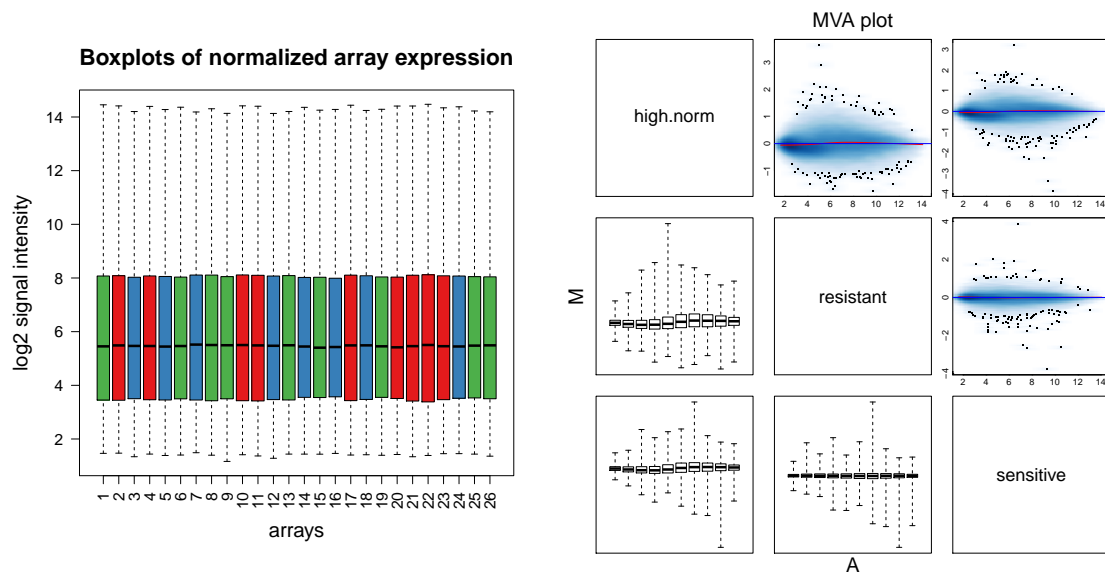
Set Parameter Values for Workflow Submission

Specifies the experimental variable to highlight in some of the qc plots \*  ▼

**Figure 6.3:** Input of registered datasets and parameters for the microarray QC workflow.

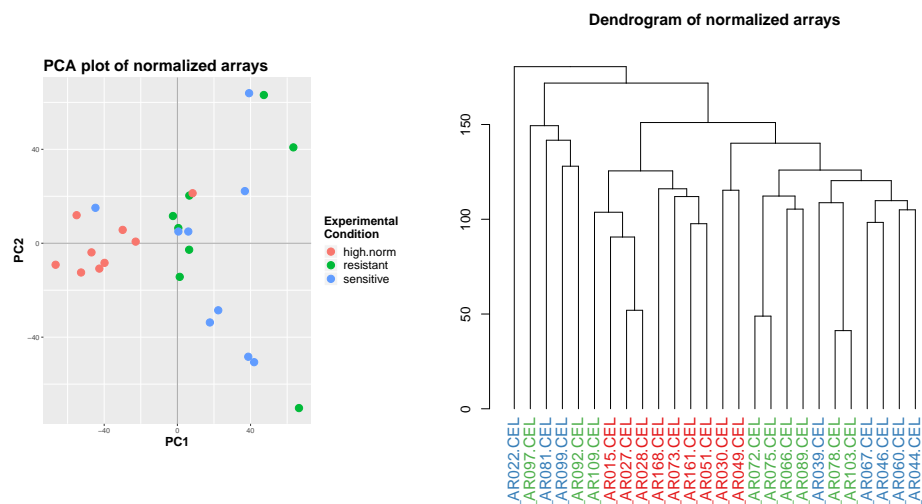
The results of our workflow suggest a successful data normalization. Figures 6.4 show that quantiles of different arrays are very similar. MA plots between different groups show that there is no general bias towards higher signal intensities in any of the groups. Several outliers can be seen, suggesting differential expression.

## 6. Using Experimental Design for Data Processing and Visualization



**Figure 6.4:** Left: Box plot of normalized microarray expression values, colored by different factor levels. Right: MA plot to discover dependence of reported differential expression between different levels on signal intensity.

Plots of the first two principal components and the dendrogram created by hierarchical clustering (Figure 6.5) suggest a similarity between samples taken of patients that show average ARU (high normal group). Interestingly, some of the aspirin-resistant and aspirin-sensitive

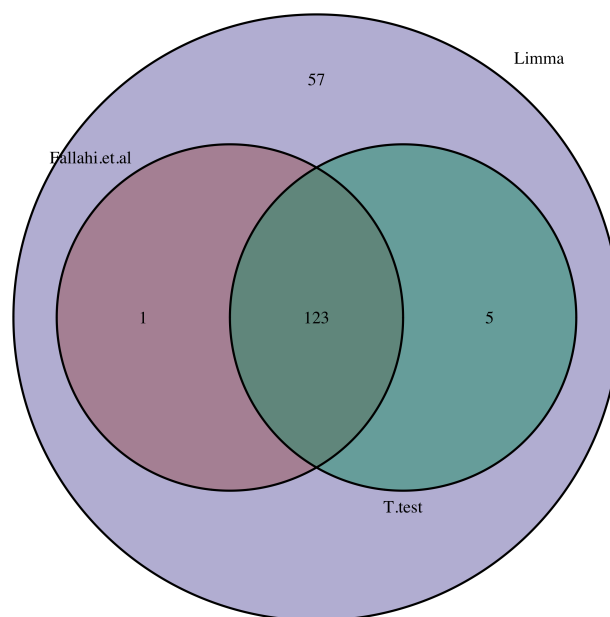


**Figure 6.5:** Sample-wise clustering colored by the different levels of aspirin effectiveness. Left: Plot of the first two principal components of the normalized expression values. Right: Hierarchical clustering of the normalized expression values.

samples cluster much closer to each other than to samples of the high normal group.

### 6.3.4 Differential Expression Analysis

Using the same thresholds as in the study by Fallahi et al.<sup>225</sup>, normal t-tests declared 128 transcripts as differentially expressed, missing one of the reported probes identified by the authors, but finding five additional transcripts. Using limma, we identified significantly more transcripts (186) as can be seen in Figure 6.6.



**Figure 6.6:** Venn diagram of differentially expressed genes according to different approaches (see Section 6.2.2 for selection criteria). Using limma resulted in the largest number (186) of reported DE transcripts, including those found by Fallahi et al.<sup>225</sup> and our own t-tests.

Limma, our t-tests and the results of Fallahi et al.<sup>225</sup> agree on the differential expression of 123 transcripts. Limma was able to identify all of the previously found transcripts. A full list of all 186 annotated transcripts including fold changes can be found in Appendix D.

Multiple testing using the Benjamini-Hochberg method with a threshold of  $\alpha = 0.05$  resulted in four significantly DE transcripts using a normal t-test. 14 additional transcripts could be identified using limma. Two groups of two transcripts had the same gene annotations, resulting in the 16 genes shown in Table 6.1.

## 6. Using Experimental Design for Data Processing and Visualization

**Table 6.1:** Significantly expressed genes found by the limma package using Benjamini-Hochberg correction. Positive fold change denotes up-regulation in the resistant group. \*Genes also significant using normal t-test with BH correction. † Mean of two significant transcripts for the same gene annotation. ‡ Affymetrix probe set ID, since no annotation was available.

Gene symbol	Log FC	Gene title
POLK*	0.63	polymerase (DNA directed) kappa
CLEC7A*	0.88 <sup>†</sup>	C-type lectin domain family 7, member A
CYP3A43*	0.50	cytochrome P450, family 3, subfamily A, polypeptide 43
HNMT*	0.62	histamine N-methyltransferase
BIN3	1.63	bridging integrator 3
HSPE1-MOB4/ MOB4	0.77	HSPE1-MOB4 readthrough, MOB family member 4, phocein HSPE1-MOB4
DPY30	0.67	dpy-30 histone methyltransferase complex regulatory subunit
222316_at <sup>‡</sup>	-0.75	—
COMMD6	0.84	COMM domain containing 6
MMP25-AS1	-0.86	MMP25 antisense RNA 1
MYOT	-0.79	myotilin
ZNF34	0.47	zinc finger protein 34
NF1	-0.71	neurofibromin 1
CNIH4	0.65 <sup>†</sup>	cornichon family AMPA receptor auxiliary protein 4
ADCY10P1	-0.75	adenylate cyclase 10 (soluble) pseudogene 1
USMG5	0.56	up-regulated during skeletal muscle growth 5 homolog (mouse)

Using limma, differential expression analysis between the aspirin-sensitive and resistant groups identified 494 transcripts with fold changes  $\geq 1.5$  and  $p \leq 0.05$ , 47 more than reported by Fallahi et al.<sup>225</sup>. Myotilin was declared as significantly down-regulated in the aspirin-resistant group. Interestingly, Myotilin has not been directly implicated in relation to blood clotting and is only known to be expressed in skeletal muscles and the heart, where it interacts with  $\alpha$ -actinin<sup>231</sup>. Nonetheless, since  $\alpha$ -actinin is known to interact with several binding-partners, further examination of the connection with myotilin might be worthwhile<sup>232</sup>.

For our second approach using normal t-tests, we selected the top 3.5% (1913) transcripts with the smallest p-values from the whole dataset. While all of the p-values for these transcripts were  $\leq 0.027$ , none of the transcripts were significantly differentially expressed after multiple testing correction, for either of the methods.

Functional analysis of the limma transcript set using DAVID<sup>230</sup> found useful gene annotations for 331 transcripts. Table 6.2 shows the 10 most significant functional annotations found for the limma results.

Comparison with functional annotation performed by Fallahi et al.<sup>225</sup> reveals that our results are highly similar: we found the four identical function annotations alternative splicing, complications of blood transfusions, coronary artery stent thrombosis, and ECT receptor inter-

**Table 6.2:** Ten most significant functional annotations provided by DAVID for the differentially expressed genes between AS and AR groups as reported by limma. Count, the percentage with respect to the provided geneset and p-values are shown.

Functional Annotation	Count	Count%	P-Value
Alternative splicing	179	54.1	$4.8 \times 10^{-4}$
Tobacco Use Disorder	72	21.8	$5.7 \times 10^{-4}$
extracellular space	37	11.2	$5.8 \times 10^{-4}$
lymphoproliferative disorders; blood transfusion complications	3	0.9	$8.0 \times 10^{-4}$
platelet degranulation	8	2.4	$1.1 \times 10^{-3}$
positive regulation of endothelial cell apoptotic process	4	1.2	$1.2 \times 10^{-3}$
coronary artery stent thrombosis	3	0.9	$1.6 \times 10^{-3}$
vaso-occlusive crisis	3	0.9	$1.6 \times 10^{-3}$
Antiphospholipid Syndrome   Arteriosclerosis Lupus Erythematosus, Systemic   Thrombosis	3	0.9	$1.6 \times 10^{-3}$
ECM-receptor interaction	7	2.1	$1.7 \times 10^{-3}$

action. Furthermore, two functional annotations that were very similar to the ones reported by Fallahi et al.<sup>225</sup> were found: platelet degranulation and thrombosis. However, we could not find significant enrichment of the functions cytoplasmic vesicle part, restenosis or different infarctions in our gene set, but instead found that 72 differentially expressed transcripts (21.8%) fit genes related to the diagnosis of Tobacco Use Disorder (TUD). Since tobacco use increases the risk of heart disease, stroke, atherosclerosis, and vascular disease<sup>233,234</sup>, we used this subset of genes to test how many of the reported genes were found in functional annotations of the other indicated categories and not linked primarily to the psychological effects of tobacco addiction. We found that of the 72 genes in this set, 59 were annotated for alternative splicing, 12 for hypertension and 10 for stroke, suggesting a relation to cardiovascular disease and the aim of the study. On the other hand, one of the inclusion criteria for patients recruited by Fallahi et al.<sup>225</sup> was risk factors, including smoking. It is conceivable that the authors did not control for this confounding factor.

Furthermore, 37 transcripts (12.2%) were annotated as relating to functions in the extracellular space, four transcripts (1.2%) were involved in the regulation of endothelial cell apoptosis and three (0.9%) to vaso-occlusive crisis, a type of thrombosis associated with sickle cell disease (SCD)<sup>235</sup>. Platelets in SCD patients are chronically activated and cause inflammation, playing a large role in this complication<sup>236</sup>.

Functional annotation results of the 3.5% of transcripts with the lowest p-value found useful annotations for 1,605 transcripts. Functional groups did not show enrichment of transcripts linked to blood clotting, platelets or related diseases. Instead, transcripts linked to mitochondrial inner membrane (67 genes, p-value  $2.6 \times 10^{-10}$ ), mRNA splicing (43 genes, p-value  $1.5 \times 10^{-8}$ ), ribonucleoproteins (50 genes, p-value  $4.0 \times 10^{-10}$ ) and cell cycle (67 genes, p-value  $7.4 \times 10^{-10}$ )

were overrepresented. This suggests that the selected, high number of genes contained too much noise, a hypothesis which is supported by the small sample size and fits the correspondingly high FDR that was predicted by OCplus. This is also reflected in the p-values of the t-test, which showed none of the genes as significant after multiple testing correction.

The R script used in the normalization and expression analyses is available as part of the results of our project on qPortal (see Appendix D). Code for our quality control workflow is available on Github.<sup>i</sup>

### 6.4 Discussion

We presented how our experimental design framework can be used to help reproduce publicly available studies, as shown by our comparison to the work performed by Fallahi et al.<sup>225</sup>. Our approach allowed the import and registration of existing ISA-Tab metadata of a study on the aspirin response signal. The aggregation graph implemented previously allows visualization of the experimental setup and sample sizes of the aspirin study.

Our implementation of a microarray QC workflow leverages the data model presented earlier to enable users to start quality control on relevant data and perform explorative data analysis based on experimental factors. Our QC analysis of the aspirin study showed successful data normalization and no biases between arrays. We noticed that hierarchical clustering and plotting of the first two principal components showed some grouping according to the different factor levels. Surprisingly, some of the low ARU (AR) samples clustered closer to the high ARU (AS) samples than to the high normal group. On the one hand, clustering based on all transcript intensities should not be over-interpreted. On the other hand, this very small group of patients might show other genetic or phenotypic similarities between groups, that Fallahi et al.<sup>225</sup> could not control for.

We performed statistical power estimation using the study's experimental design. Results showed that, based on the sample sizes of the two groups of aspirin-sensitive and aspirin-resistant samples, we could expect only low sensitivity, while false discovery rate was relatively high, especially for transcripts with low fold changes.

Using the same thresholds for  $\alpha$  and fold change, and the standard implementation of the t-test in R, we could confirm the large majority of transcripts found by Fallahi et al.<sup>225</sup> to be DE between the aspirin-resistant group and all other samples. One of the transcripts found to be DE in the previous study was not detected by our t-test but was found by limma. In addition, both of our methods reported five additional transcripts as differentially expressed. As discussed in Section 2.3 varying software or operating system versions used in the analysis may lead to rounding errors. The proximity of the p-values and fold change in question to the cut-off values of 0.001 and 1.5, respectively, indicate this as a reasonable cause. Except for this

---

<sup>i</sup>[https://github.com/qbicsoftware/qbic-wf-microarray\\_qc-old](https://github.com/qbicsoftware/qbic-wf-microarray_qc-old)

outlier, we could reproduce this part of the study with high sensitivity. The results show that our analysis using the limma R package reported 57 unique transcripts as differentially expressed. This can be attributed to its more complex approach of using linear models and an empirical Bayes method to moderate the standard errors of estimated fold changes, resulting in improved power<sup>106</sup>. Interestingly, the limma results declared myotilin as significantly down-regulated in the aspirin-resistant group, a connection that may encourage further study due to its relation to well-known binding partners<sup>231,232</sup>.

We used limma for the second part of our analysis, focusing on differential expression between sensitive and resistant samples. While we found overall more genes to be differentially expressed, we found very similar functional annotations corresponding to platelet function and related disease states. For example, vaso-occlusive crisis is a type of thrombosis associated with sickle cell disease (SCD)<sup>235</sup>. Platelets in SCD patients are chronically activated and cause inflammation, playing a large role in this complication<sup>236</sup>. As aspirin inhibits platelet function, these annotations fit well to the scope of genetic mechanisms for aspirin resistance or sensitivity.

Functional annotation also suggested 72 genes linked to tobacco use disorder. We showed that at least some of these genes are in fact known to be involved in cardiovascular diseases. However, this connection could also be explained by insufficient correction for the confounding factor of smoking by the authors of the original study.

We also used the top 3.5% transcripts with lowest p-values for functional annotation. This approach did not reveal relevant functional groups, which potentially points towards too many genes being included, leading to noise. This explanation is supported by both the high FDR that was predicted by OCplus, as well as the results of the corrected t-test. On the other hand, most corrections for multiple testing are very strict by design and might not be a fitting benchmark for the method proposed by Pawitan et al.<sup>128</sup>. Nonetheless, to further test the benefits of this approach, larger datasets or datasets with larger fold changes are needed. Selecting the top differentially expressed genes without performing multiple testing may seem unorthodox. However, it can be compared to the common methods of alleviating the multiple testing problem by previously filtering out genes with low variability or fold change. Additionally, the need to control for false discovery rate is always dependent on the aim of a study. If few promising candidate genes need to be found, strict filtering with adjusted p-values is necessary. For some approaches, for example, the training of a machine learning algorithm, researchers may want to use a larger set of genes. Here, a less strict approach might provide benefits as validation using additional data needs to be done in any case.

In summary, we could recreate many of the results reported by Fallahi et al.<sup>225</sup> using their methods and described threshold values. As expected, following the analysis protocol as closely as possible has a large impact on the results. However, despite using the same analysis and parameters, there was a very slight disagreement regarding which transcripts were reported as differentially expressed. The latter part of our analysis shows that the use of publically

available online tools can be problematic for reproducibility: DAVID provides a very large number of options, especially concerning the annotation databases. Fallahi et al.<sup>225</sup> did not record the version or parameters used and we could only guess that they used the GAD disease database from their results. Versioning in this context is also important since many of the source databases used by tools like DAVID are updated concurrently to the algorithm itself. Nonetheless, we were able to replicate many of the functional groups found in the earlier study, even when using the larger transcript set reported by limma. This suggests that many of the platelet-related functional annotations may actually be significant.

We did not use a qPortal workflow for our differential expression analysis for several reasons. First, we wanted to use various analysis methods and types of thresholds, including top-ranking genes as well as p-values. Secondly, our experience from related studies has taught us that the annotation of microarray data is notoriously difficult, as different platform versions cannot be automatically obtained from the same source. Here, container solutions, as discussed in previous chapters, may be beneficial for solving versioning problems. Microarray data also lends itself very well to interactive analysis, as none of the analysis steps after normalization typically take longer than a few seconds. This allows for fast troubleshooting and testing of different approaches, which is why interactive analysis tools like Mayday have been developed<sup>237</sup>. Nonetheless, qPortal does provide the necessary metadata and computational framework as a foundation to include workflows for the most often used platforms and methods. Current development is focused on providing these tools.

## Chapter 7

# Conclusion and Outlook

In recent years, vast amounts of data have been generated in genomics, transcriptomics, proteomics, and other fields of high-throughput biology and medicine. Despite the tremendous opportunities that have been opened up by the increasingly fast and cheap measurement of big biological data, researchers have identified many obstacles that stand in the way of efficient and accurate exploitation of these data. The reproducibility crisis has shown that many studies cannot be reliably reproduced. The main reasons for these difficulties are missing or incorrect metadata annotation on the one hand and insufficient statistical planning of experiments on the other.

Especially due to the fact that big biological data is often associated with large and varied metadata annotation, specialized models, concepts, and tools are required to collect metadata and other information about experimental designs. The shift from localized data analysis to the cloud and different online platforms has led to the creation of various science gateways. Since their main purpose is to allow users to analyze different types of omics data, these platforms often focus on the annotation of workflow parameters. In contrast, our first contribution focuses on the tracking of the biological provenance of samples using a tier-based data model. The Experimental Design Wizard portlet helps researchers create experiments and samples as part of a multi-factorial research design. It has been used extensively as part of the qPortal platform. Creation of large studies is simplified by allowing the import of sample tables or the open metadata standard ISA-Tab. Adding further metadata annotation is possible via the upload of spreadsheets. While we already cover two distinct use cases in the creation of experimental designs and the entry of additional measurements for existing samples, we are working on further modularizing these two aspects of project creation. Support for further metadata formats like the ones used by ENA, SRA or PRIDE are under development. We have shown the extensibility of our approach for peptidomics, ligandomics, and other experiment types and support for additional study and data types is in development. The correct way of performing integrative data analysis based on different omics types is still a major topic of research that has not been conclusively solved. Therefore, scientists are in need of new analysis approaches and pipelines. Our software and data model builds a foundation that enables researchers to successfully plan multi-omics studies, integrate raw data and develop integrative analysis pipelines that take into account the study design.

While spreadsheet-based formats have strengthened the connection between lab-scientists and data science, they lose some of their usefulness for large or even medium-sized projects. Current methods are not sufficient to allow scientists to check the validity of the vast amounts of metadata they enter. We have discovered that many publically available studies show inconsistencies between the experimental design reported in ISA-Tab and their publications. Our contribution to resolving this problem is an interactive visualization of large studies based on the aggregation of similar samples into a graph showing the biological provenance. Here, we make use of our data model and experimental designs to visualize study factors and number of replicates or other interesting aspects of a study. Through our support of the open format ISA-Tab, public studies can be easily explored. While we make our approach available via a stand-alone application, the real strength of this visualization is highlighted in combination with additional data sources as they are available in qPortal instances. Imported studies can be checked for consistency before they are registered in the system. For existing studies, missing data can be visualized and the respective samples listed. Since our approach is not strictly tied to the aggregation of study factors or species, we are working on the inclusion of more customization options for users. Visualization of aggregation graphs shows promise for future applications like the interactive creation or editing of study designs.

To focus on the second cause of the reproducibility crisis — insufficient experimental design and missing statistical planning — we connected our data model to an interactive method for sample size and power estimation in differential gene expression analysis. While a number of methods for power analysis exist in this field, many are limited to command line tools or statistical programming languages and none are connected to metadata and data storage. With our approach, we enable the easy use of pilot data to predict sample size or statistical power of an experiment, potentially saving costs and helping researchers to create reproducible studies. While we focus on data of microarrays and RNA-Sequencing, our portlet is extendable for other methods and data types, given that tools can be included in a Singularity container or even directly into the portlet. This modularity is useful, as there is still no agreement on the best methods to estimate the required sample size or the power of studies using RNA-Seq data. An important point is that our focus in these cases was on studies comparing two levels of a factor. Our model for experimental designs supports multi-factorial research designs and factors with more than two levels. To take full advantage of these designs, additional methods are needed for power prediction and analysis of the experiments created with our approaches. Analysis of variance (ANOVA) is one classic example that can be used to compare three or more group means. For microarrays, the analysis package Limma provides methods to analyze complex experimental setups. Power analysis can also be a topic of interest for multi-omics experiments, where taking into account the interaction between different omics layers is important. Methods that use the knowledge gained through the use of one technology could potentially increase

---

statistical power when other parts of the data is insufficient to draw statistically significant conclusions.

The application of our work to a real study shows that we can easily import and analyze well-annotated studies. In the course of this analysis, we leveraged the functionality of qPortal and our experimental design model in order to create a workflow for automated normalization and quality control of microarray data, further demonstrating the advantages of our approach. Utilization of our portlet for statistical power estimation showed that sample sizes were most certainly not large enough to control for both type I and type II errors. In general, the comparison of differential gene expression analysis based on the approaches used by the original study was a success. However, despite the authors providing good annotation for their data, we could identify obstacles when reproducing their study. Especially when online tools are used, the provenance chain from raw data to results can be broken, if no version information is provided. If these tools are based on third-party databases that are frequently updated between versions, the problem can be exacerbated. Guidelines are needed so providers of online services prominently offer versioning and parameter information of their algorithms, and researchers need to be made aware that they should provide this information upon publication.

Existing guidelines, as defined by the FAIR standard, already try to alleviate the causes of the reproducibility crisis. Our approaches provide FAIR data in the following ways: the Experimental Design Wizard enables input, upload and update of metadata for large studies, connected to the powerful FAIR data management system openBIS. This lays the foundation to make data findable by its metadata. For statistical power estimations and analysis workflows, we store parameters and results, guaranteeing data provenance and reproducibility. As part of qPortal, metadata imported via our tools, and the attached data are stored and transferred securely and using accessible protocols. Interoperability is achieved by using controlled vocabularies for suitable types of metadata, often matched to public ontologies of the respective biomedical fields. In addition, we support open standards like ISA-Tab and are working on export functionality for multiple formats. Last but not least, we aid in the reusability of our results and tools by providing the source code and versions of portlets, containers, and workflows using Github and Maven.

In conclusion, we are confident that our efforts will contribute to the advancement of various fields of biological high-throughput analysis.



# Bibliography

- [1] Doug Howe, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, et al. Big data: The future of biocuration. *Nature*, 455(7209):47, 2008. 1
- [2] Rasko Leinonen, Ruth Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, Nadeem Faruque, Neil Goodgame, Richard Gibson, et al. The european nucleotide archive. *Nucleic Acids Research*, 39(suppl\_1):D28–D31, 2010. 61, 70
- [3] Charles E Cook, Mary Todd Bergman, Robert D Finn, Guy Cochrane, Ewan Birney, and Rolf Apweiler. The European Bioinformatics Institute in 2016: data growth and integration. *Nucleic Acids Research*, 44(D1):D20–D26, 2015. 1
- [4] Jakob Lovén, David A Orlando, Alla A Sigova, Charles Y Lin, Peter B Rahl, Christopher B Burge, David L Levens, Tong Ihn Lee, and Richard A Young. Revisiting global gene expression analysis. *Cell*, 151(3):476–482, 2012. 1
- [5] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003. 1, 8
- [6] Katja Dettmer, Pavel A Aronov, and Bruce D Hammock. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1):51–78, 2007. 1
- [7] Nina Hillen and Stefan Stevanovic. Contribution of mass spectrometry-based proteomics to immunology. *Expert Review of Proteomics*, 3(6):653–664, 2006. 1, 9
- [8] Anna Reustle, Moreno Di Marco, Carolin Meyerhoff, Annika Nelde, Juliane S Walz, Stefan Winter, Siahei Kandabarau, Florian Büttner, Mathias Haag, Linus Backert, et al. Integrative-omics and hla-ligandomics analysis to identify novel drug targets for ccrcc immunotherapy. *Genome Medicine*, 12(1):1–24, 2020. 1, 9
- [9] Murtha Baca. *Introduction to metadata*. Getty Publications, 2016. 1
- [10] Alvis Brazma, Pascal Hingamp, John Quackenbush, Gavin Sherlock, Paul Spellman, Chris Stoeckert, John Aach, Wilhelm Ansorge, Catherine A Ball, Helen C Causton, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics*, 29(4):365–371, 2001. 2, 23
- [11] Alvis Brazma. Minimum information about a microarray experiment (MIAME)—successes, failures, challenges. *The Scientific World Journal*, 9:420–423, 2009. 2, 23, 29
- [12] Paul T Spellman, Michael Miller, Jason Stewart, Charles Troup, Ugis Sarkans, Steve Chervitz, Derek Bernhart, Gavin Sherlock, Catherine Ball, Marc Lepage, et al. Design and implementation of

- microarray gene expression markup language (MAGE-ML). *Genome Biology*, 3(9):research0046–1, 2002. 2, 23
- [13] Chris F Taylor, Norman W Paton, Kathryn S Lilley, Pierre-Alain Binz, Randall K Julian, Andrew R Jones, Weimin Zhu, Rolf Apweiler, Ruedi Aebersold, Eric W Deutsch, et al. The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology*, 25(8):887–893, 2007. 2, 23, 29
- [14] Juan Antonio Vizcaíno, Richard G Côté, Attila Csordas, José A Dianes, Antonio Fabregat, Joseph M Foster, Johannes Griss, Emanuele Alpi, Melih Birim, Javier Contell, et al. The PRoteomics IDentifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Research*, 41(D1):D1063–D1069, 2012. 2
- [15] Susanna-Assunta Sansone, Philippe Rocca-Serra, Marco Brandizi, Alvis Brazma, Dawn Field, Jennifer Fostel, Andrew G Garrow, Jack Gilbert, Federico Goodsaid, Nigel Hardy, et al. The first RSBI (ISA-TAB) workshop: “can a simple format work for complex studies?”. *OMICS A Journal of Integrative Biology*, 12(2):143–149, 2008. 2, 23, 24
- [16] Susanna-Assunta Sansone, Philippe Rocca-Serra, Dawn Field, Eamonn Maguire, Chris Taylor, Oliver Hofmann, Hong Fang, Steffen Neumann, Weida Tong, Linda Amaral-Zettler, et al. Toward interoperable bioscience data. *Nature Genetics*, 44(2):121, 2012. 2, 24, 29
- [17] Philippe Rocca-Serra, Marco Brandizi, Eamonn Maguire, Nataliya Sklyar, Chris Taylor, Kimberly Begley, Dawn Field, Stephen Harris, Winston Hide, Oliver Hofmann, et al. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, 26(18):2354–2356, 2010. 2, 24, 29, 45, 62, 70
- [18] Alejandra González-Beltrán, Eamonn Maguire, Susanna-Assunta Sansone, and Philippe Rocca-Serra. linkedISA: semantic representation of ISA-Tab experimental metadata. *BMC Bioinformatics*, 15(14):S4, 2014. 2, 24, 29, 62
- [19] Konstantinos CM Siontis, Nikolaos A Patsopoulos, and John PA Ioannidis. Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. *European Journal of Human Genetics*, 18(7):832–837, 2010. 2
- [20] Joshua Carp. The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage*, 63(1):289–300, 2012. 57
- [21] Jeffery T Leek and Roger D Peng. What is the question? *Science*, 347(6228):1314–1315, 2015.
- [22] Jeffrey T Leek and Roger D Peng. Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences*, 112(6):1645–1646, 2015. 2, 22, 30
- [23] Mark Liberman. Replicability vs. reproducibility—or is it the other way around. *The Language Log*, 2015. 2, 22

- 
- [24] Anton Nekrutenko and James Taylor. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, 13(9):667–672, 2012. 2, 57
- [25] Daniel Garijo, Sarah Kinnings, Li Xie, Lei Xie, Yinliang Zhang, Philip E Bourne, and Yolanda Gil. Quantifying reproducibility in computational biology: the case of the tuberculosis drugome. *PLoS ONE*, 8(11):e80278, 2013. 2, 57
- [26] Stephen R Piccolo and Michael B Frampton. Tools and techniques for computational reproducibility. *GigaScience*, 5(1):30, 2016. 2
- [27] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 2016. 2, 24, 25, 58
- [28] Florian Prinz, Thomas Schlange, and Khusru Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10(9):712–712, 2011. 2
- [29] Francis S Collins and Lawrence A Tabak. NIH plans to enhance reproducibility. *Nature*, 505(7485):612, 2014. 2
- [30] Qunhua Li, James B Brown, Haiyan Huang, Peter J Bickel, et al. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3):1752–1779, 2011. 2
- [31] Jeffrey T Leek and Leah R Jager. Is most published research really false? *Annual Review of Statistics and Its Application*, 4:109–122, 2017. 2
- [32] Shannon E Ellis and Jeffrey T Leek. How to share data for collaboration. *The American Statistician*, 72(1):53–57, 2018. 2, 22
- [33] Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005. 3
- [34] Joel N Hirschhorn, Kirk Lohmueller, Edward Byrne, and Kurt Hirschhorn. A comprehensive review of genetic association studies. *Genetics in Medicine*, 4(2):45, 2002. 3
- [35] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorf, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747, 2009. 3
- [36] Ronald Aylmer Fisher. *The design of experiments*. Oliver And Boyd; Edinburgh; London, 1937. 3, 15
- [37] Sonia Tarazona, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. Differential expression in RNA-seq: a matter of depth. *Genome Research*, pages gr-124321, 2011. 3, 82

- [38] Kasper D Hansen, Zhijin Wu, Rafael A Irizarry, and Jeffrey T Leek. Sequencing technology does not eliminate biological variability. *Nature Biotechnology*, 29(7):572, 2011. 82
- [39] Yuwen Liu, Jie Zhou, and Kevin P White. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 30(3):301–304, 2013. 3, 82
- [40] Tim F Rayner, Philippe Rocca-Serra, Paul T Spellman, Helen C Causton, Anna Farne, Ele Holloway, Rafael A Irizarry, Junmin Liu, Donald S Maier, Michael Miller, et al. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, 7(1):489, 2006. 3, 23
- [41] International Human Genome Sequencing Consortium et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860, 2001. 7
- [42] Rodger Staden. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, 6(7):2601–2610, 1979. 7
- [43] Al Edwards, Hartmut Voss, Peter Rice, Andrew Civitello, Josef Stegemann, Christian Schwager, Juergen Zimmermann, Holger Erfle, C Thomas Caskey, and Wilhelm Ansorge. Automated DNA sequencing of the human HPRT locus. *Genomics*, 6(4):593–608, 1990. 7
- [44] James L Weber and Eugene W Myers. Human whole-genome shotgun sequencing. *Genome Research*, 7(5):401–409, 1997. 7
- [45] Michael L Metzker. Sequencing technologies—the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010. 7
- [46] James M Heather and Benjamin Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, 2016. 7
- [47] Afshin Ahmadian, Maria Ehn, and Sophia Hober. Pyrosequencing: History, biochemistry and future. *Clinica Chimica Acta*, 363(1-2):83–94, 2006. 7
- [48] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008. 7
- [49] Karl V Voelkerding, Shale A Dames, and Jacob D Durtschi. Next-generation sequencing: from basic research to diagnostics. *Clinical Chemistry*, 55(4):641–658, 2009. 7
- [50] Jonathan M Rothberg, Wolfgang Hinz, Todd M Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H Leamon, Kim Johnson, Mark J Milgrew, Matthew Edwards, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, 2011. 8
- [51] Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1):1–11, 2016. 8

- 
- [52] Elaine R Mardis. DNA sequencing technologies: 2006–2016. *Nature Protocols*, 12(2):213, 2017. 8, 10
- [53] HPJ Buermans and JT Den Dunnen. Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1842(10):1932–1941, 2014. 8, 10
- [54] Peter J Park. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009. 8
- [55] Xu Li, Wenqi Wang, and Junjie Chen. Recent progress in mass spectrometry proteomics for biomedical research. *Science China Life Sciences*, 60(10):1093–1113, 2017. 8, 9
- [56] Merriam-Webster. Spectrometer. *Merriam-Webster.com dictionary*. URL <https://www.merriam-webster.com/dictionary/spectrometer>. 8
- [57] Matthias Mann, Nils A Kulak, Nagarjuna Nagaraj, and Jürgen Cox. The coming age of complete, accurate, and ubiquitous proteomes. *Molecular Cell*, 49(4):583–590, 2013. 8
- [58] Nasrin Mirsaleh-Kohan, Wesley D Robertson, and Robert N Compton. Electron ionization time-of-flight mass spectrometry: Historical review and current applications. *Mass Spectrometry Reviews*, 27(3):237–285, 2008. 8
- [59] Christian Jurinke, Paul Oeth, and Dirk van den Boom. MALDI-TOF mass spectrometry. *Molecular Biotechnology*, 26(2):147–163, 2004. 8
- [60] Joseph A Loo. Studying noncovalent protein complexes by electrospray ionization mass spectrometry. *Mass Spectrometry Reviews*, 16(1):1–23, 1997. 8
- [61] Franz Hillenkamp, Michael Karas, Ronald C Beavis, and Brian T Chait. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Analytical Chemistry*, 63(24):1193A–1203A, 1991. 8
- [62] John B Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craig M Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, 1989. 8
- [63] John B Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craige M Whitehouse. Electrospray ionization—principles and practice. *Mass Spectrometry Reviews*, 9(1):37–70, 1990. 8
- [64] Anas El-Aneel, Aljandro Cohen, and Joseph Banoub. Mass spectrometry, review of the basics: electrospray, MALDI, and commonly used mass analyzers. *Applied Spectroscopy Reviews*, 44(3):210–230, 2009. 8
- [65] BA Mamyrin. Time-of-flight mass spectrometry (concepts, achievements, and prospects). *International Journal of Mass Spectrometry*, 206(3):251–266, 2001. 8
- [66] IUPAC. mass-to-charge ratio. *Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*, 2019. URL <https://doi.org/10.1351/goldbook.M03752>. 8

- [67] Philip E Miller and M Bonner Denton. The quadrupole mass filter: basic operating concepts. *Journal of Chemical Education*, 63(7):617, 1986. 9
- [68] Edmond De Hoffmann and Vincent Stroobant. *Mass spectrometry: principles and applications*. John Wiley & Sons, 2007. 9
- [69] Dirk A Wolters, Michael P Washburn, and John R Yates. An automated multidimensional protein identification technology for shotgun proteomics. *Analytical Chemistry*, 73(23):5683–5690, 2001. 9
- [70] Guodong Chen and Birendra N Pramanik. Application of LC/MS to proteomics studies: current status and future prospects. *Drug Discovery Today*, 14(9-10):465–471, 2009. 9
- [71] Alexey I Nesvizhskii. Protein identification by tandem mass spectrometry and sequence database searching. *Mass Spectrometry Data Analysis in Proteomics*, pages 87–119, 2007. 9
- [72] Jeremy E Melanson, Kenneth A Chisholm, and Devanand M Pinto. Targeted comparative proteomics by liquid chromatography/matrix-assisted laser desorption/ionization triple-quadrupole mass spectrometry. *Rapid Communications in Mass Spectrometry*, 20(5):904–910, 2006. 9
- [73] Hubert Chassaigne, Jørgen V Nørgaard, and Arjon J van Hengel. Proteomics-based approach to detect and identify major allergens in processed peanuts by capillary LC-Q-TOF (MS/MS). *Journal of Agricultural and Food Chemistry*, 55(11):4461–4473, 2007. 9
- [74] Sander R Piersma, Marc O Warmoes, Meike de Wit, Inge de Reus, Jaco C Knol, and Connie R Jiménez. Whole gel processing procedure for GeLC-MS/MS based proteomics. *Proteome Science*, 11(1):1–9, 2013. 9
- [75] Patric Hörth, Christine A Miller, Tobias Preckel, and Christian Wenz. Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis. *Molecular & Cellular Proteomics*, 5(10):1968–1974, 2006. 9
- [76] Shao-En Ong, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics*, 1(5):376–386, 2002. 9
- [77] Matthias Mann. Functional and quantitative proteomics using SILAC. *Nature Reviews Molecular Cell Biology*, 7(12):952–958, 2006. 9
- [78] Philip L Ross, Yulin N Huang, Jason N Marchese, Brian Williamson, Kenneth Parker, Stephen Hattan, Nikita Khainovski, Sasi Pillai, Subhakar Dey, Scott Daniels, et al. Multiplexed protein quantitation in *saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics*, 3(12):1154–1169, 2004. 9
- [79] Sebastian Wiese, Kai A Reidegeld, Helmut E Meyer, and Bettina Warscheid. Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. *Proteomics*, 7(3):340–350, 2007. 9

- 
- [80] Andrew Thompson, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Schmidt, Thomas Neumann, and Christian Hamon. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, 75(8):1895–1904, 2003. 9
- [81] Paul J Boersema, Reinout Raijmakers, Simone Lemeer, Shabaz Mohammed, and Albert JR Heck. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nature Protocols*, 4(4):484–494, 2009. 9
- [82] Friedrich Lottspeich and Josef Kellermann. ICPL labeling strategies for proteome research. In *Gel-Free Proteomics*, pages 55–64. Springer, 2011. 9
- [83] Eric S Witze, William M Old, Katheryn A Resing, and Natalie G Ahn. Mapping protein post-translational modifications with mass spectrometry. *Nature Methods*, 4(10):798–806, 2007. 9
- [84] Ruedi Aebersold and Matthias Mann. Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620):347–355, 2016. 9
- [85] Aleksandra Nita-Lazar, Hideshiro Saito-Benz, and Forest M White. Quantitative phosphoproteomics by mass spectrometry: past, present, and future. *Proteomics*, 8(21):4433–4443, 2008. 9
- [86] Boris Macek, Matthias Mann, and Jesper V Olsen. Global and site-specific quantitative phosphoproteomics: principles and applications. *Annual Review of Pharmacology and Toxicology*, 49:199–221, 2009. 9
- [87] Sheng Pan, Ru Chen, Ruedi Aebersold, and Teresa A Brentnall. Mass spectrometry based glycoproteomics—from a proteomics perspective. *Molecular & Cellular Proteomics*, 10(1):R110–003251, 2011. 9
- [88] Leila Afjehi-Sadat and Benjamin A Garcia. Comprehending dynamic protein methylation with mass spectrometry. *Current Opinion in Chemical Biology*, 17(1):12–19, 2013. 9
- [89] Donald S Kirkpatrick, Carilee Denison, and Steven P Gygi. Weighing in on ubiquitin: the expanding role of mass-spectrometry-based proteomics. *Nature Cell Biology*, 7(8):750–757, 2005. 9
- [90] Nikolai Mischerikow and Albert JR Heck. Targeted large-scale analysis of protein acetylation. *Proteomics*, 11(4):571–589, 2011. 9
- [91] James Robinson, Jason A Halliwell, James D Hayhurst, Paul Flicek, Peter Parham, and Steven GE Marsh. The ipd and imgt/hla database: allele variant databases. *Nucleic Acids Research*, 43(D1):D423–D431, 2015. 9
- [92] Ana Marcu, Leon Bichmann, Leon Kuchenbecker, Linus Backert, Daniel J Kowalewski, Lena Katharina Freudenmann, Markus W Löffler, Maren Lübke, Juliane S Walz, Julia Velz, et al. The HLA Ligand Atlas. A resource of natural HLA ligands presented on benign tissues. *BioRxiv*, page 778944, 2019. 9

- [93] Bernard Marr. Big data: The 5 Vs everyone must know. *LinkedIn Pulse*, 6, 2014. 9
- [94] Illumina, Inc. NovaSeq System Specifications, 2021. URL <https://www.illumina.com/systems/sequencing-platforms/novaseq/specifications.html>. 10
- [95] Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big data: Astronomical or Genomical? *PLoS Biology*, 13(7):e1002195, 2015. 10
- [96] Edward J Fox, Kate S Reid-Bayliss, Mary J Emond, and Lawrence A Loeb. Accuracy of next generation sequencing platforms. *Journal of Next Generation Sequencing & Applications*, 1, 2014. 10
- [97] Heather D VanGuilder, Kent E Vrana, and Willard M Freeman. Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques*, 44(5):619–626, 2008. 10
- [98] Andrew I Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences*, 101(16):6062–6067, 2004. 10
- [99] Kassie S Manning and Thomas A Cooper. The roles of RNA processing in translating genotype to phenotype. *Nature Reviews Molecular Cell Biology*, 18(2):102–114, 2017. 10
- [100] Christopher Lee and Meenakshi Roy. Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biology*, 5(7):1–4, 2004. 11
- [101] Emily Clough and Tanya Barrett. The gene expression omnibus database. In *Statistical Genomics*, pages 93–110. Springer, 2016. 11
- [102] Yuk Fai Leung and Duccio Cavalieri. Fundamentals of cDNA microarray data analysis. *TRENDS in Genetics*, 19(11):649–659, 2003. 11, 13
- [103] Yongjun Chu and David R Corey. RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Therapeutics*, 22(4):271–274, 2012. 11
- [104] Randall K Saiki, David H Gelfand, Susanne Stoffel, Stephen J Scharf, Russell Higuchi, Glenn T Horn, Kary B Mullis, and Henry A Erlich. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239(4839):487–491, 1988. 12
- [105] Joseph DeRisi, L Penland, ML Bittner, PS Meltzer, M Ray, Y Chen, YA Su, and JM Trent. Use of a cDNA microarray to analyse gene expression. *Nature Genetics*, 14:457–460, 1996. 12
- [106] Gordon K Smyth and Terry Speed. Normalization of cDNA microarray data. *Methods*, 31(4):265–273, 2003. 13, 86, 95
- [107] Illumina, Inc. Multiplexed Sequencing with the Illumina Genome Analyzer System. 2008. URL [https://www.illumina.com/documents/products/datasheets/datasheet\\_sequencing\\_multiplex.pdf](https://www.illumina.com/documents/products/datasheets/datasheet_sequencing_multiplex.pdf). 13

- 
- [108] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):1–13, 2013. 13
- [109] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013. 13
- [110] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621, 2008. 14
- [111] Günter P Wagner, Koryu Kin, and Vincent J Lynch. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, 131(4):281–285, 2012. 14
- [112] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. 14, 20, 21
- [113] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010. 14, 20, 21
- [114] David J Glass. *Experimental design for biologists*. Number QH323. 5 G52. Cold Spring Harbor Laboratory Press Cold Spring Harbour, NY, USA, 2014. 14
- [115] R Fisher. Introduction to “The arrangement of field experiments”. *J Minist Agric G B*, 33:503–13, 1926. 15
- [116] Thomas Eden and Ronald A Fisher. Studies in crop variation: VI. Experiments on the response of the potato to potash and nitrogen. *The Journal of Agricultural Science*, 19(2):201–213, 1929. 15
- [117] Andreas Friedrich, Luis de la Garza, Oliver Kohlbacher, and Sven Nahnsen. Interactive Visualization for Large-Scale Multi-factorial Research Designs. In *International Conference on Data Integration in the Life Sciences*, pages 75–84. Springer, 2018. 15
- [118] Graeme Ruxton and Nick Colegrave. *Experimental design for the life sciences*. Oxford University Press, 2011. 15, 16, 17
- [119] Veronica Czitrom. One-factor-at-a-time versus designed experiments. *The American Statistician*, 53(2):126–131, 1999. 16
- [120] Crispin Y Jordan. Population sampling affects pseudoreplication. *PLoS Biology*, 16(10):e2007054, 2018. 17
- [121] Xiangqin Cui and Gary A Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, 4(4):1–10, 2003. 18
- [122] Chris Cheadle, Marquis P Vawter, William J Freed, and Kevin G Becker. Analysis of microarray data using Z score transformation. *The Journal of Molecular Diagnostics*, 5(2):73–81, 2003. 18

- [123] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006. 18
- [124] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, 57(1):289–300, 1995. 18, 87
- [125] Carlo E Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. *Studi in onore del professore salvatore ortu carboni*, pages 13–60, 1935. 18
- [126] Eric W Weisstein. Bonferroni Correction – From MathWorld – A Wolfram Web Resource, 2004. URL <https://mathworld.wolfram.com/BonferroniCorrection.html>. 18
- [127] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70, 1979. 18
- [128] Yudi Pawitan, Stefan Michiels, Serge Koscielny, Arief Gusnanto, and Alexander Ploner. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, 21(13):3017–3024, 2005. 18, 74, 81, 87, 95
- [129] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297, 2012. 21, 75, 82
- [130] Yi-Hui Zhou, Kai Xia, and Fred A Wright. A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics*, 27(19):2672–2678, 2011. 21
- [131] Hao Wu, Chi Wang, and Zhijin Wu. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, 14(2):232–243, 2012. 21
- [132] David Roxbee Cox and Nancy Reid. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(1):1–18, 1987. 21
- [133] Jon F Claerbout and Martin Karrenbach. Electronic documents give reproducible research a new meaning. In *SEG Technical Program Expanded Abstracts 1992*, pages 601–604. Society of Exploration Geophysicists, 1992. 22
- [134] Roger D Peng, Francesca Dominici, and Scott L Zeger. Reproducible epidemiologic research. *American Journal of Epidemiology*, 163(9):783–789, 2006. 22
- [135] Siri Carpenter. Psychology’s bold initiative. *Science*, 335(6076):1558–1561, 2012. ISSN 0036-8075. doi: 10.1126/science.335.6076.1558. URL <https://science.sciencemag.org/content/335/6076/1558>. 22
- [136] Sohyun Hwang, Eiru Kim, Insuk Lee, and Edward M Marcotte. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, 5:17875, 2015. 22

- 
- [137] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, 2017. 22, 27
- [138] Felix W Frueh. Impact of microarray data quality on genomic data submissions to the FDA. *Nature Biotechnology*, 24(9):1105, 2006. 23
- [139] Weida Tong, Stephen C Harris, Hong Fang, Leming Shi, Roger Perkins, Federico Goodsaid, and Felix W Frueh. An integrated bioinformatics infrastructure essential for advancing pharmacogenomics and personalized medicine in the context of the FDA’s Critical Path Initiative. *Drug Discovery Today: Technologies*, 4(1):3–8, 2007. 23
- [140] Marcia McNutt. Journals unite for reproducibility. *Science*, 346(6210):679–679, 2014. ISSN 0036-8075. doi: 10.1126/science.aaa1724. URL <https://science.sciencemag.org/content/346/6210/679>. 23
- [141] Juan A Vizcaíno, Eric W Deutsch, Rui Wang, Attila Csordas, Florian Reisinger, Daniel Ríos, José A Dianes, Zhi Sun, Terry Farrah, Nuno Bandeira, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology*, 32(3):223–226, 2014. 23, 61, 70
- [142] Sandra Orchard, Lukasz Salwinski, Samuel Kerrien, Luisa Montecchi-Palazzi, Matthias Oesterheld, Volker Stümpflen, Arnaud Ceol, Andrew Chatr-Aryamontri, John Armstrong, Peter Woollard, et al. The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature Biotechnology*, 25(8):894, 2007. 23
- [143] Daniel Kolarich, Erdmann Rapp, Weston B Struwe, Stuart M Haslam, Joseph Zaia, Ryan McBride, Sanjay Agravat, Matthew P Campbell, Masaki Kato, Rene Ranzinger, et al. The minimum information required for a glycomics experiment (MIRAGE) project: improving the standards for reporting mass-spectrometry-based glycoanalytic data. *Molecular & Cellular Proteomics*, 12(4):991–995, 2013. 23
- [144] Chris F Taylor, Dawn Field, Susanna-Assunta Sansone, Jan Aerts, Rolf Apweiler, Michael Ashburner, Catherine A Ball, Pierre-Alain Binz, Molly Bogue, Tim Booth, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology*, 26(8):889, 2008. 23
- [145] Lennart Martens, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander, Jim Shofstahl, Wilfred H Tang, Andreas Römpf, Steffen Neumann, Angel D Pizarro, et al. mzML—a community standard for mass spectrometry data. *Molecular & Cellular Proteomics*, 10(1):R110–000133, 2011. 23
- [146] Johannes Griss, Andrew R Jones, Timo Sachsenberg, Mathias Walzer, Laurent Gatto, Jürgen Hartler, Gerhard G Thallinger, Reza M Salek, Christoph Steinbeck, Nadin Neuhauser, et al. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Molecular & Cellular Proteomics*, 13(10):2765–2775, 2014. 23

- [147] Andrew R Jones, Michael Miller, Ruedi Aebersold, Rolf Apweiler, Catherine A Ball, Alvis Brazma, James DeGreef, Nigel Hardy, Henning Hermjakob, Simon J Hubbard, et al. The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. *Nature Biotechnology*, 25(10):1127, 2007. 24
- [148] Mark D Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, and Michel Dumontier. A design framework and exemplar metrics for FAIRness. *Scientific Data*, 5(1):1–4, 2018. 25
- [149] Mark D Wilkinson, Michel Dumontier, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Mario Prieto, Dominique Batista, Peter McQuilton, Tobias Kuhn, Philippe Rocca-Serra, Mercè Crosas, et al. Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Scientific Data*, 6(1):1–12, 2019.
- [150] Daniel C Berrios, Afshin Beheshti, and Sylvain V Costes. FAIRness and usability for open-access omics data systems. In *AMIA Annual Symposium Proceedings*, volume 2018, page 232. American Medical Informatics Association, 2018. 25
- [151] Peter Kunszt, Lorenz Blum, Béla Hullár, Emanuel Schmid, Adam Srebniak, Witold Wolski, Bernd Rinn, Franz-Josef Elmer, Chandrasekhar Ramakrishnan, Andreas Quandt, et al. iPortal: the swiss grid proteomics portal: Requirements and new features based on experience and usability considerations. *Concurrency and Computation: Practice and Experience*, 27(2):433–445, 2015. 26, 29, 57, 61
- [152] Jeremy Goecks, Anton Nekrutenko, and James Taylor. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, 2010. 26, 61
- [153] Michael Reich, Ted Liefeld, Joshua Gould, Jim Lerner, Pablo Tamayo, and Jill P Mesirov. GenePattern 2.0. *Nature Genetics*, 38(5):500, 2006. 26, 29, 57, 61
- [154] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404, 2012. 26
- [155] Shayan Shahand, Ammar Benabdelkader, Jordi Huguet, Mohammad Mahdi Jaghoori, Mark Santcroos, Mostapha al Mourabit, Paul FC Groot, Matthan WA Caan, Antoine HC van Kampen, and Sílvia D Olabarriaga. A Data-Centric Science Gateway for Computational Neuroscience. In *IWSG*, 2013. 26
- [156] Mark A Miller, Wayne Pfeiffer, and Terri Schwartz. The CIPRES science gateway: a community resource for phylogenetic analyses. In *Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery*, page 41. ACM, 2011. 26
- [157] Jelle Scholtalbers, Jasmin Rößler, Patrick Sorn, Jos de Graaf, Valesca Boisguérin, John Castle, and Ugur Sahin. Galaxy LIMS for next-generation sequencing. *Bioinformatics*, 29(9):1233–1234, 2013. 26

- 
- [158] Ravi K Madduri, Dinanath Sulakhe, Lukasz Lacinski, Bo Liu, Alex Rodriguez, Kyle Chard, Utpal J Dave, and Ian T Foster. Experiences building Globus Genomics: a next-generation sequencing analysis service using Galaxy, Globus, and Amazon Web Services. *Concurrency and Computation: Practice and Experience*, 26(13):2266–2279, 2014. 26
- [159] Marko Grönroos. *The Book of Vaadin: Vaadin 7 Edition*, volume 4th Revision. Vaadin Ltd, 2014. 26, 43
- [160] Marc Fleury and Francisco Reverbel. The JBoss extensible server. In *Proceedings of the ACM/I-FIP/USENIX 2003 International Conference on Middleware*, pages 344–373. Springer-Verlag New York, Inc., 2003. 26
- [161] Alejandro Abdelnur and Stefan Hepper. JSR 168: Portlet specification. *Java Specification Requests, Java Community Process, Sun Microsystems and IBM*, 15, 2003. 26
- [162] Adaptive Computing. Moab HPC suite, 2015. 26
- [163] Garrick Staples. TORQUE resource manager. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, page 8. ACM, 2006. 26
- [164] Andy B Yoo, Morris A Jette, and Mark Grondona. Slurm: Simple linux utility for resource management. In *Workshop on Job Scheduling Strategies for Parallel Processing*, pages 44–60. Springer, 2003. 26
- [165] Matthew Portnoy. *Virtualization essentials*, volume 19. John Wiley & Sons, 2012. 26
- [166] Stephen Soltesz, Herbert Pötzl, Marc E Fiuczynski, Andy Bavier, and Larry Peterson. Container-based operating system virtualization: a scalable, high-performance alternative to hypervisors. In *ACM SIGOPS Operating Systems Review*, volume 41, pages 275–287. ACM, 2007. 27
- [167] Dirk Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014(239):2, 2014. 27
- [168] Gregory M Kurtzer, Vanessa Sochat, and Michael W Bauer. Singularity: Scientific containers for mobility of compute. *PloS ONE*, 12(5):e0177459, 2017. 27, 76
- [169] Theo Combe, Antony Martin, and Roberto Di Pietro. To Docker or not to Docker: A security perspective. *IEEE Cloud Computing*, 3(5):54–62, 2016. 27
- [170] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452, 2016. 29
- [171] IEEE Standards Association. IEEE Recommended Practice for Software Requirements Specifications, 1998. URL <https://standards.ieee.org/standard/830-1998.html>. 31
- [172] Angela Bauch, Izabela Adamczyk, Piotr Buczek, Franz-Josef Elmer, Kaloyan Enimanev, Pawel Glyzewski, Manuel Kohler, Tomasz Pylak, Andreas Quandt, Chandrasekhar Ramakrishnan, et al. openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics*, 12(1):468, 2011. 31, 46, 61

- [173] Thomas Allweyer. *BPMN 2.0: introduction to the standard for business process modeling*. BoD–Books on Demand, 2016. 35
- [174] Jakob Nielsen. *Usability engineering*. Elsevier, 1994. 37, 58
- [175] Stuart K Card. *The psychology of human-computer interaction*. CRC Press, 2017. 37
- [176] Allen Newell. *Unified theories of cognition*. Harvard University Press, 1994. 38, 58
- [177] Scott Federhen. The NCBI taxonomy database. *Nucleic Acids Research*, 40(D1):D136–D143, 2011. 39
- [178] Alejandra Gonzalez-Beltran and David Johnson. ISA Data Model, 2018. URL <https://isa-specs.readthedocs.io/en/latest/isamodel.html>. 41
- [179] Inc Liferay. Liferay Portal, 2018. URL <http://www.liferay.com/>. 43
- [180] The Apache Software Foundation. Apache Tomcat 7, 2018. URL <http://tomcat.apache.org/tomcat-7.0-doc/>. 43
- [181] Jani Hursti. Single sign-on. In *Proc. Helsinki University of Technology Seminar on Network Security*, 1997. 43
- [182] Riccardo Murri, Peter Z Kunszt, Sergio Maffioletti, and Valery Tschopp. GridCertLib: a single sign-on solution for Grid web applications and portals. *Journal of Grid Computing*, 9(4):441–453, 2011. 43
- [183] Atlassian Corporation Plc. Atlassian Crowd, 2018. URL <http://www.atlassian.com/software/crowd/>. 43
- [184] Frederic P Miller, Agnes F Vandome, and John McBrewster. Apache Maven. 2010. 45
- [185] Andrew Tridgell, Paul Mackerras, et al. The rsync algorithm. 1996. 46
- [186] Christopher Mohr, Andreas Friedrich, David Wojnar, Erhan Kenar, Aydin Can Polatkan, Marius Cosmin Codrea, Stefan Czettel, Oliver Kohlbacher, and Sven Nahnsen. qPortal: A platform for data-driven biomedical research. *PLoS ONE*, 13(1):e0191603, 2018. 51, 59, 86, 89
- [187] Stuart J Johnston. Wizards’ make microsoft applications smarter. *InfoWorld*, 13(31):6, 1991. 57
- [188] Martin Fowler. *Patterns of Enterprise Application Architecture*. Addison-Wesley, 2012. 58
- [189] Alejandra Gonzalez-Beltran, Eamonn Maguire, Pavlos Georgiou, Susanna-Assunta Sansone, and Philippe Rocca-Serra. Bio-GraphIn: a graph-based, integrative and semantically-enabled repository for life science experimental data. *EMBNET. Journal*, 19(B):pp–46, 2013. 62, 70
- [190] Andreas Friedrich, Erhan Kenar, Oliver Kohlbacher, and Sven Nahnsen. Intuitive web-based experimental design for high-throughput biomedical data. *BioMed Research International*, 2015, 2015. 62

- 
- [191] Kenneth Haug, Reza M Salek, Pablo Conesa, Janna Hastings, Paula de Matos, Mark Rijnbeek, Tejasvi Mahendraker, Mark Williams, Steffen Neumann, Philippe Rocca-Serra, et al. Metabo-Lights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research*, 41(D1):D781–D786, 2012. 64, 70
- [192] Santosh Lamichhane, Linda Ahonen, Thomas Sparholt Dyrland, Esko Kemppainen, Heli Siljander, Heikki Hyoty, Jorma Ilonen, Jorma Toppari, Riitta Veijola, Tuulia Hyotylainen, et al. Dynamics of Plasma Lipidome in Progression to Islet Autoimmunity and Type 1 Diabetes: Type 1 Diabetes Prediction and Prevention Study (DIPP). *bioRxiv*, page 294033, 2018. 64
- [193] Chris Pettitt. dagre - Graph layout for JavaScript. <https://github.com/dagrejs/dagre>, 2014. 65
- [194] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D<sup>3</sup> data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011. 65
- [195] Yuanyuan Tian, Richard A Hankins, and Jignesh M Patel. Efficient aggregation for graph summarization. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 567–580. ACM, 2008. 70
- [196] Steven Noel and Sushil Jajodia. Managing attack graph complexity through visual hierarchical aggregation. In *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, pages 109–118. ACM, 2004. 70
- [197] Andrei Z Broder, Ronny Lempel, Farzin Maghoul, and Jan Pedersen. Efficient PageRank approximation via graph aggregation. *Information Retrieval*, 9(2):123–138, 2006. 70
- [198] Natasha A Karp and Kathryn S Lilley. Design and analysis issues in quantitative proteomics studies. *Proteomics*, 7(S1):42–50, 2007. 73, 81
- [199] Ann L Oberg and Olga Vitek. Statistical design of quantitative mass spectrometry-based proteomic experiments. *Journal of Proteome Research*, 8(5):2144–2156, 2009.
- [200] Luis Valledor and Jesús Jorrín. Back to the basics: maximizing the information obtained by quantitative two dimensional gel electrophoresis analyses by an appropriate experimental design and statistical analyses. *Journal of Proteomics*, 74(1):1–18, 2011. 73, 81
- [201] Michael Borenstein. *Power and Precision*, volume 1. Taylor & Francis, 2001. 73, 81
- [202] Russel V Lenth. Java applets for power and sample size [computer software], 2006-9. URL <http://www.stat.uiowa.edu/~rlenth/Power>. 73, 81
- [203] Meena Choi, Ching-Yun Chang, Timothy Clough, Daniel Broudy, Trevor Killeen, Brendan MacLean, and Olga Vitek. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*, 30(17):2524–2526, 2014. 73, 81
- [204] Hannes L Röst, Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aichele, Sandro Andreotti, Hans-Christian Ehrlich, Petra Gutenbrunner, Erhan Kenar, et al. OpenMS: a

- flexible open-source software platform for mass spectrometry data analysis. *Nature Methods*, 13(9):741, 2016. 73
- [205] Mei-Ling Ting Lee and George Alex Whitmore. Power and sample size for DNA microarray studies. *Statistics in Medicine*, 21(23):3543–3570, 2002. 73, 81
- [206] Caimiao Wei, Jiangning Li, and Roger E Bumgarner. Sample size for detecting differentially expressed genes in microarray experiments. *BMC Genomics*, 5(1):87, 2004.
- [207] Kevin Dobbin and Richard Simon. Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, 6(1):27–38, 2005. 73
- [208] Travers Ching, Sijia Huang, and Lana X Garmire. Power analysis and sample size estimation for RNA-Seq differential expression. *RNA*, 2014. 73
- [209] Huaieu Luo, Juntao Li, Burton Kuan Hui Chia, Paul Robson, and Niranjana Nagarajan. The importance of study design for detecting differentially abundant features in high-throughput experiments. *Genome Biology*, 15(12):1–17, 2014. 73
- [210] Mei-Ling Ting Lee and Weiliang Qiu. SPCalc: A web-based calculator for sample size and power calculations in micro-array studies. *Bioinformatics*, 2006. 74
- [211] Yan Guo, Shilin Zhao, Chung-I Li, Quanhu Sheng, and Yu Shyr. RNAseqPS: a web tool for estimating sample size and power for RNAseq experiment. *Cancer Informatics*, 13:CIN–S17688, 2014. 74
- [212] Michele A Busby, Chip Stewart, Chase A Miller, Krzysztof R Grzeda, and Gabor T Marth. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics*, 29(5):656–657, 2013. 74, 81
- [213] Shilin Zhao, Chung-I Li, Yan Guo, Quanhu Sheng, and Yu Shyr. RnaSeqSampleSize: real data based sample size estimation for RNA sequencing. *BMC Bioinformatics*, 19(1):191, 2018. 74, 81
- [214] Cancer Genome Atlas Research Network et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061, 2008. 75
- [215] Chung-I Li, Pei-Fang Su, and Yu Shyr. Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data. *BMC Bioinformatics*, 14(1):357, 2013. 75
- [216] Steven N Hart, Terry M Therneau, Yuji Zhang, Gregory A Poland, and Jean-Pierre Kocher. Calculating sample size estimates for RNA sequencing data. *Journal of Computational Biology*, 20(12):970–978, 2013. 75, 78
- [217] Yunshun Chen, Davis McCarthy, Matthew Ritchie, Mark Robinson, and Gordon Smyth. edgeR: differential expression analysis of digital gene expression data – User’s Guide, 2018. URL <https://www.bioconductor.org/packages/devel/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf>. 75, 78
- [218] Arho Virkki. Rvaadin, 2017. URL <https://github.com/avirkki/Rvaadin>. 75

- 
- [219] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 2008. 82
- [220] Colin Baigent, Lisa Blackwell, Rory Collins, Jonathan Emberson, Jon Godwin, Richard Peto, Julie Buring, Charles Hennekens, Patricia Kearney, Tom Meade, et al. Aspirin in the primary and secondary prevention of vascular disease: collaborative meta-analysis of individual participant data from randomised trials., 2009. 85
- [221] Sean L Zheng and Alistair J Roddick. Association of aspirin use for primary prevention with cardiovascular events and bleeding events: a systematic review and meta-analysis. *Jama*, 321(3):277–287, 2019. 85
- [222] George Krasopoulos, Stephanie J Brister, W Scott Beattie, and Michael R Buchanan. Aspirin “resistance” and risk of cardiovascular morbidity: systematic review and meta-analysis. *BMJ*, 336(7637):195–198, 2008. 85
- [223] Thomas H Wang, Deepak L Bhatt, and Eric J Topol. Aspirin and clopidogrel resistance: an emerging clinical entity. *European Heart Journal*, 27(6):647–654, 2005. 85
- [224] Michael B Soyka, Kaspar Rufibach, Alex Huber, and David Holzmann. Is severe epistaxis associated with acetylsalicylic acid intake? *The Laryngoscope*, 120(1):200–207, 2010. 85
- [225] Payam Fallahi, Richard Katz, Ian Toma, Ranyang Li, Jonathan Reiner, Kiersten VanHouten, Larry Carpio, Lorraine Marshall, Yi Lian, Sujata Bupp, et al. Aspirin insensitive thrombophilia: transcript profiling of blood identifies platelet abnormalities and HLA restriction. *Gene*, 520(2):131–138, 2013. 86, 91, 92, 93, 94, 95, 96, 128
- [226] Inc. Accriva Diagnostics. Verify Now Reference Guide, 2016. URL [http://www.accriva.com/uploads/literature/mvn0005\\_verifynow\\_pocket\\_guide\\_01.pdf](http://www.accriva.com/uploads/literature/mvn0005_verifynow_pocket_guide_01.pdf). 86
- [227] Tal Galili. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, 31(22):3718–3720, 2015. 86
- [228] Peter Kacsuk, Zoltan Farkas, Miklos Kozlovsky, Gabor Hermann, Akos Balasko, Krisztian Karoczkai, and Istvan Marton. WS-PGRADE/gUSE generic DCI gateway framework for a large variety of user communities. *Journal of Grid Computing*, 10(4):601–630, 2012. 86
- [229] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005. 86
- [230] Da Wei Huang, Brad T Sherman, Xin Zheng, Jun Yang, Tomozumi Imamichi, Robert Stephens, and Richard A Lempicki. Extracting biological meaning from large gene lists with DAVID. *Current Protocols in Bioinformatics*, 27(1):13–11, 2009. 87, 92
- [231] Paula Salmikangas, Peter FM van der Ven, Maciej Lalowski, Anu Taivainen, Fang Zhao, Heli Suila, Rolf Schröder, Pekka Lappalainen, Dieter O Fürst, and Olli Carpén. Myotilin, the limb-girdle muscular dystrophy 1A (LGMD1A) protein, cross-links actin filaments and controls sarcomere assembly. *Human Molecular Genetics*, 12(2):189–203, 2003. 92, 95

- [232] Carol A Otey and Olli Carpen.  $\alpha$ -actinin revisited: A fresh look at an old player. *Cell Motility and the Cytoskeleton*, 58(2):104–111, 2004. 92, 95
- [233] Reena S Shah and John W Cole. Smoking and stroke: the more you smoke the more you stroke. *Expert Review of Cardiovascular Therapy*, 8(7):917–932, 2010. 93
- [234] George Howard, Lynne E Wagenknecht, Gregory L Burke, Ana Diez-Roux, Gregory W Evans, Paul McGovern, F Javier Nieto, Grethe S Tell, ARIC investigators, et al. Cigarette smoking and progression of atherosclerosis: The Atherosclerosis Risk in Communities (ARIC) Study. *Jama*, 279(2):119–124, 1998. 93
- [235] Deepa Manwani and Paul S Frenette. Vaso-occlusion in sickle cell disease: pathophysiology and novel targeted therapies. *Blood*, 122(24):3892–3898, 2013. 93, 95
- [236] Erica Sparkenbaugh and Rafal Pawlinski. Interplay between coagulation and vascular inflammation in sickle cell disease. *British Journal of Haematology*, 162(1):3–14, 2013. 93, 95
- [237] Janko Dietzsch, Nils Gehlenborg, and Kay Nieselt. Mayday-a microarray data analysis workbench. *Bioinformatics*, 22(8):1010–1012, 2006. 96





## Appendix A: Abbreviations

---

API	<i>Application Programming Interface</i>
bp	<i>base pair</i>
cDNA	<i>complementary DNA</i>
CTD	<i>Common tool descriptor</i>
DE	<i>Differentially expressed/differential expression</i>
ENA	<i>European Nucleotide Archive</i>
ESI	<i>electrospray ionization</i>
ETL	<i>Extract, transform, load</i>
FDR	<i>False discovery rate</i>
FN(R)	<i>False negative (rate)</i>
FPKM	<i>Fragments per kilobase per million mapped reads</i>
(G)UI	<i>(Graphical) user interface</i>
HLA	<i>human leukocyte antigen</i>
iTRAQ	<i>isobaric Tags for Relative and Absolute Quantitation</i>
I/O	<i>Input/Output</i>
LC(-MS)	<i>Liquid chromatography(–mass spectrometry)</i>
LDAP	<i>Lightweight directory access protocol</i>
LIMS	<i>Laboratory information management system</i>
MALDI	<i>matrix-assisted laser desorption/ionization</i>
MS	<i>Mass spectrometry</i>
MS/MS	<i>tandem mass spectrometry</i>
m/z ratio	<i>mass-to-charge ratio</i>
NGS	<i>Next-generation sequencing</i>
OBI	<i>Ontology for Biomedical Investigations</i>
openBIS	<i>Open Source Biology Information System</i>

## Abbreviations

---

OS	<i>Operating system</i>
PTM	<i>posttranslational modification(s)</i>
QC	<i>Quality control</i>
(Q)TOF	<i>(quadrupole) time-of-flight</i>
RIN	<i>RNA integrity number</i>
RNA-Seq	<i>RNA sequencing</i>
RPKM	<i>Reads per kilobase per million mapped reads</i>
SILAC	<i>stable isotope labeling by amino acids in cell culture</i>
SSH	<i>Secure Shell</i>
SSO	<i>Single sign-on</i>
TMT	<i>Tandem Mass Tag</i>
TP(R)	<i>True positive (rate)</i>
VM	<i>Virtual machine</i>

---

**Appendix B: Visual Exploration of Experimental Designs**

**Table of MetaboLights studies used for validation**

**Table B.1:** MetaboLights studies used in validation of our graph aggregation tool for experimental design visualization. Study identifiers, description, and our assessment of reproducibility due to the shared information and its accuracy compared with the respective publication is shown.

Identifier	Description	Assessment
MTBLS618	Metabolomics on tomato plants infected with blight	accurate metadata
MTBLS619	Response to dietary carbohydrates in seabass tissue	missing species replicates
MTBLS622	Wheat and aphid metabolism	biological replicates missing
MTBLS640	Metabolome in Parkinson's mouse model	factor undefined, no replicates
MTBLS654	Role of fermented fish in kimchi fermentation	accurate metadata
MTBLS669	Effects of starvation in breast cancer model	accurate metadata
MTBLS674	Metabolome in Parkinson's mouse model	factor undefined, no replicates
MTBLS687	Linking root exudates to functional plant traits	3 replicates missing
MTBLS750	Effects of Lipin1 deficiency	accurate metadata
MTBLS780	Multi-omics study on <i>S. cerevisiae</i> strain diversity	ISA-Tab contains additional strains

## **Appendix C: Interactive Sample Size Calculation for Differential Gene Expression Experiments**

### Code used for Statistical Power Estimations

```
1 #Calculate FDR for FC=1 and different sample sizes of a Microarray
   ↪ experiment
2 library(OCPlus)
3 samplesize(p0 = 0.95, D=1, crit.style = c('top_percentage'), crit=0.05)
```

## **Appendix D: Using Experimental Design for Data Processing and Visualization**

## Supplementary material

Supplementary material is available via qPortal.<sup>i</sup> using user credentials *qbcdemo01* and password *demo*.

## Table of all differentially expressed transcripts found

**Table D.1:** Significantly expressed transcripts found by the Limma package filtered by minimum fold change of 1.5 and a threshold of  $\alpha = 0.001$ . Positive fold change denotes up-regulation in the resistant group. \*Probes also found to be DE in the study by Fallahi et al.<sup>225</sup>.

Probe Set ID	Gene Symbol	log FC
1564052_at*	TREML4	-1.673
1557582_at*	BIN3	1.629
215101_s_at	CXCL5	-1.508
241092_at	—	-1.472
1555520_at	PTCH1	-1.230
1565716_at*	FUS	-1.204
220117_at*	ZNF385D	-1.162
220374_at*	KLHL28	-1.125
238913_at*	—	-1.090
241339_at*	—	-1.073
243169_at	—	1.059
1569676_at	TOR1AIP2	-1.051
1569685_at*	COX10	-1.046
202110_at*	COX7B	1.027
239501_at*	—	-0.999
215663_at*	MBNL1	-0.992
209773_s_at	RRM2	-0.960
217329_x_at*	—	0.958
222791_at*	RSBN1	-0.957
241254_at	—	-0.953
215708_s_at	PRIM2	-0.951
244349_at*	—	-0.934
227088_at	PDE5A	-0.923
229095_s_at*	LIMS3-LOC440895	0.920
201742_x_at*	SRSF1	-0.905
219659_at	ATP8A2	-0.904
222777_s_at*	WHSC1	-0.903
1555756_a_at*	CLEC7A	0.898
239799_at	LINC00476	-0.897
241732_at*	—	-0.892
236669_at*	SDCBP2-AS1	0.891
1559648_at	LINC00892	-0.883
239384_at*	SRSF1	-0.881
202635_s_at*	POLR2K	0.880
208782_at	FSTL1	-0.880
238231_at*	NFYC	-0.877
226289_at*	CAPRN1	-0.874
1557315_a_at*	—	-0.873
204724_s_at	COL9A3	0.873
236513_at	PRELID2	-0.871
1562528_at	—	-0.869
1554406_a_at*	CLEC7A	0.864
227449_at*	EPHA4	-0.859
240054_at*	MMP25-AS1	-0.857
233725_at*	—	0.857
1570352_at*	ATM	-0.841
225312_at*	COMMD6	0.839
238573_at*	OTUD7B	-0.836
229773_at*	SNAP23	-0.836
223566_s_at*	BCOR	-0.834
213280_at	RAP1GAP2	-0.833
244332_at*	—	-0.831
1552646_at	IL11RA	-0.826
1559614_at*	FLJ38773	0.822
236511_at*	—	-0.817

---

<sup>i</sup>[https://portal.qbic.uni-tuebingen.de/portal/web/qbic/browser#!project//PUBLIC\\_PROJECTS/QHASP](https://portal.qbic.uni-tuebingen.de/portal/web/qbic/browser#!project//PUBLIC_PROJECTS/QHASP)

Probe Set ID	Gene Symbol	log FC
240703_s_at	HERC1	-0.816
1560171_at*	—	-0.812
233228_at*	—	-0.808
228542_at*	MRS2	-0.808
213846_at	COX7C	0.806
219728_at*	MYOT	-0.793
227016_at*	ERICH1	0.791
239426_at*	SLC2A8	-0.785
213285_at*	TMEM30B	-0.784
207550_at	MPL	-0.783
237389_at	—	-0.780
203960_s_at*	HSPB11	0.777
241786_at*	—	-0.776
233925_at	—	-0.772
222875_at	DHX33	-0.772
1563958_at*	—	-0.771
217535_at	—	0.770
225195_at*	DPH3	0.769
1554309_at*	EIF4G3	-0.769
234989_at*	NEAT1	-0.769
202918_s_at*	HSPE1-MOB4 /// MOB4	0.767
228033_at	E2F7	-0.765
215898_at*	TLL5	-0.753
222316_at*	—	-0.752
1557300_s_at	—	-0.752
233959_at*	ADCY10P1	-0.750
1556758_at*	FAM208B	-0.750
228053_s_at*	TOMM5	0.749
229391_s_at	FAM26F	0.749
218049_s_at	MRPL13	0.745
1557749_at*	EHBP1L1	0.742
238578_at*	TMEM182	-0.741
226130_at*	RPS16	-0.738
211732_x_at*	HNMT	0.735
242968_at*	—	-0.731
200061_s_at*	RPS24	0.730
228984_at	CARNS1	-0.728
233099_at*	—	-0.727
201754_at	COX6C	0.727
202634_at*	POLR2K	0.721
212678_at*	NF1	-0.715
1560622_at	—	-0.715
223797_at*	PRO2852	-0.710
222575_at*	SETD5	-0.708
232264_at*	—	-0.706
232406_at*	LOC105372526	0.701
213767_at*	KSR1	0.699
230171_at	LOC105377348	0.698
218728_s_at*	CNIH4	0.691
223062_s_at	PSAT1	-0.691
1566129_at	LIMS1	0.691
220969_s_at*	—	-0.690
202984_s_at	BAG5	-0.690
1558183_at*	ZNF17	-0.689
227461_at*	STON2	-0.685
243996_at	NPR2	-0.682
206572_x_at*	ZNF85	0.681
227156_at*	CASK	-0.681
232676_x_at*	MYEF2	-0.676
239760_at	—	-0.675
225127_at*	TMEM181	-0.674
214224_s_at*	PIN4	0.673
211973_at	NUDT3	0.670
214657_s_at	NEAT1	-0.670
224129_s_at*	DPY30	0.669
236207_at*	SSFA2	0.669
219794_at*	VPS53	0.666
1558369_at*	MPHOSPH9	-0.664
203703_s_at	TLL4	-0.663
1568732_at	—	0.663
229374_at	EPHA4	-0.661
1570048_at	DNAJC24	-0.659
243003_at*	—	-0.658
236165_at*	MSL3	-0.654
209000_s_at*	SEPT8	-0.654
1561596_at*	—	0.654
222771_s_at	MYEF2	-0.651
227350_at*	HELLS	-0.648
217682_at*	C16orf72	-0.648
1560680_at*	—	-0.647
202278_s_at	SPTLC1	-0.645
206007_at	PRG4	-0.645
235433_at*	APOOL	-0.640
228556_at*	YTHDC1	0.639

Probe Set ID	Gene Symbol	log FC
214717_at*	PKI55	-0.638
217773_s_at	NDUFA4	0.638
225036_at*	TOMM5	0.636
244190_at*	THAP5	0.635
223260_s_at*	POLK	0.634
241250_at*	—	-0.631
205644_s_at*	SNRPG	0.628
204112_s_at*	HNMT	0.623
214242_at*	MAN1A2	-0.623
1562903_at	FAM86B3P	0.622
224511_s_at*	TXNDC17	0.620
235847_at	—	-0.615
1554986_a_at*	SNX19	0.614
204375_at	CLSTN3	-0.614
200963_x_at*	RPL31	0.612
218351_at*	COMMD8	0.611
215407_s_at*	ASTN2	-0.611
213822_s_at*	UBE3B	0.610
229319_at*	—	-0.607
1557126_a_at*	PLD1	-0.606
221380_at	—	0.606
1552633_at*	ZNF101	0.605
1553530_a_at*	ITGB1 /// ITGB1P1	-0.605
221434_s_at*	SLIRP	0.605
227840_at	C2orf76	0.605
244132_x_at*	ZNF518A	-0.604
223996_s_at	MRPL30	0.604
223993_s_at*	CNIH4	0.603
203781_at*	MRPL33	0.602
1556633_at*	C1orf204	-0.602
204992_s_at*	PFN2	-0.601
222044_at	PCIF1	0.601
1552306_at	ALG10	-0.601
243129_at*	CXorf40A /// CXorf40B	-0.600
203521_s_at*	ZNF318	-0.598
243279_at*	—	0.598
240893_at	LOC101928317	0.597
241932_at*	—	-0.595
218309_at*	CAMK2N1	-0.594
243903_at*	—	0.593
233818_at	LTN1	-0.591
243993_at*	—	-0.588
233476_at*	—	0.588
217998_at	PHLDA1	-0.587
214747_at	ZBED4	-0.586
242676_at	NDUFV2-AS1	0.586
225080_at	MYO1C	-0.585