Challenges in Integration and Analysis of High-Dimensional Biological Data: Cases from Environmental and Health Research

by

Mariam Reyad Rizkallah

A thesis submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
in Data Engineering

Approved Dissertation Committee

Prof. Dr. Adalbert F.X. Wilhelm
Jacobs University Bremen

Prof. Dr. Frank Oliver Glöckner
Jacobs University Bremen

Prof. Dr. Iris Pigeot
Leibniz Institute for Prevention
Research and Epidemiology – BIPS

Date of Defense: January 31, 2022

Computer Science & Electrical Engineering

Satutory Declaration

Family Name, Given/First Name	Rizkallah, Mariam Reyad
Matriculation number	20331431
Type of thesis	PhD

English: Declaration of Authorship

I hereby declare that the thesis submitted was created and written solely by myself without any external support. Any sources, direct or indirect, are marked as such. I am aware of the fact that the contents of the thesis in digital form may be revised with regard to usage of unauthorized aid as well as whether the whole or parts of it may be identified as plagiarism. I do agree my work to be entered into a database for it to be compared with existing sources, where it will remain in order to enable further comparisons with future theses. This does not grant any rights of reproduction and usage, however.

The Thesis has been written independently and has not been submitted at any other university for the conferral of a PhD degree; neither has the thesis been previously published in full.

German: Erklärung der Autorenschaft (Urheberschaft)

Ich erkläre hiermit, dass die vorliegende Arbeit ohne fremde Hilfe ausschließlich von mir erstellt und geschrieben worden ist. Jedwede verwendeten Quellen, direkter oder indirekter Art, sind als solche kenntlich gemacht worden. Mir ist die Tatsache bewusst, dass der Inhalt der Thesis in digitaler Form geprüft werden kann im Hinblick darauf, ob es sich ganz oder in Teilen um ein Plagiat handelt. Ich bin damit einverstanden, dass meine Arbeit in einer Datenbank eingegeben werden kann, um mit bereits bestehenden Quellen verglichen zu werden und dort auch verbleibt, um mit zukïunftigen Arbeiten verglichen werden zu können. Dies berechtigt jedoch nicht zur Verwendung oder Vervielfältigung.

Diese Arbeit wurde in der vorliegenden Form weder einer anderen Prüfungsbehörde vorgelegt noch wurde das Gesamtdokument bisher veröffentlicht.

November 15, 2021

_____

Date, Signature

Acknowledgements

Since I was a little girl, I was fascinated by the power of diagnosis and treatment. I dreamed, as did my pharmacist parents, with a world of sustained access to the right treatment for the right persons. In college, I was introduced to the magical world of bioinformatics, and the knowledge that genetic data hold and tell when properly interrogated using the right procedures. I was fortunate to conduct computational research at the elite institutions I was privileged to join.

I am particularly grateful to Jacobs University Bremen - Data Engineering Program for giving me the opportunity to follow my dreams, granting me the best education and training. During my PhD, I was learning something new every day. My thesis project would have never been possible without the unique expertise and innovative (super)-vision of my thesis advisors. I deeply thank my supervisor, Prof. Dr. Adalbert Wilhelm for his endless support, time and guidance. It was a pleasure being supervised by you, navigating data engineering with your comprehensive focused guidance. I thank my co-supervisor Prof. Dr. Frank Oliver Glöckner for his guidance, support, patience, time, and fruitful discussions.

I am eternally immensely grateful to my direct supervisor, Prof. Dr. Iris Pigeot, the Director of the Leibniz Institute for Prevention Research and Epidemiology – BIPS. It was an honor working on my thesis project under your supervision. It was beyond my dreams to be supervised by the first-time awardee of "Outstanding Doctoral Supervision" prize. I experienced firsthand how you give amble space for creativity for your students, so we can in turn give our best. You also have a deep consideration of the individual challenges of a PhD project, and you calmly walk us through them. Thank you for bringing my thesis project to light.

I am exceptionally very grateful to each and every person who helped me throughout my career path up to this very moment. From BIPS, I thank my former boss Dr. Ronja Foraita, and former colleagues Dr. Marvin N. Wright and Louis Dijkstra, Prof. Dr. Vanessa Didelez, and my fellow PhD students for their support and helpful discussions.

My graduate studies' journey was toughest on my family. I am immensely grateful to my mother, father and brother for supporting me all the way throughout my graduate studies. I thank my husband, Dimitar, for his endless support, for fruitful discussions, and for believing in me. I thank my little one, Damian, for taking it easy on me and bearing with me in this adventure.

Last but not least, I thank Al-Alfi Foundation for Human and Social Development for their generous financial and moral support for my PhD research stay. I was very fortunate to be part of your Leaders Program. I am grateful to your sustained financial support even during the toughest times of the Egyptian economy.

JACOBS UNIVERSITY BREMEN

Abstract

Computer Science & Electrical Engineering

Doctor of Philosophy

Challenges in Integration and Analysis of High-Dimensional Biological Data: Cases from Environmental and Health Research

by Mariam Reyad Rizkallah

Biological data represent a large, challenging sector of data engineering applications, where being "drowned in data and starved for information" could cost human lives. Biological data are typically complex and poorly standardized. Moreover, high value, rapid growth in volume and advances in acquisition technologies characterize modern environmental and health research data, humbling the classical practices for data transformation and analytics. Furthermore, data in biology make more sense when integrated with usually different data types, or data from different sources or even fields. In addition, the uniqueness of each case and research question call for a deep understanding of data life cycle and for customized solutions. Having a large volume and value, and being produced at a high velocity in a large variety, biological data encourage the investigation of scalable workflows to automate acquisition and integration, closing the gaps in optimizing analytics specially for heterogeneous data.

This thesis aims at exploring and optimizing the state-of-the-art methods for heterogeneous data integration and analysis, of sequence and non-sequence-based data, by identifying four areas of application concerning primary and secondary data from environmental and health research. It presents four challenges in data preparation and transformation for variable selection, and accompanying case studies. Particularly, the thesis investigates knowledge extraction from primary inherently high-dimensional marine sequence data, scalability in handling secondary photosynthetic sequence data, integration and statistical modeling of secondary high-dimensional relational health care claims data for adverse drug event prediction, and integration of heterogeneous primary epidemiological data for childhood obesity investigation. The thesis highlights the importance of data model development for data transformation and integration, and the role of scalable analytics in the foreseen increase in data dimensions.

Contents

List of Figures

List of Tables

List of Abbreviations

24-HDR	24-Hour Dietary Recall
ADE	Adverse Drug Event
API	Application Programming Interface
ARTP	The Adaptive Rank Truncated Product
ATC	Anatomical Therapeutic Chemical Classification System
AWS	Amazon Web Services
BF	Block Forest
BIPS	The Leibniz Institute for Prevention Research and Epidemiology – BIPS GmbH
BLAST	Basic Local Alignment Search Tool
BMI	Body Mass Index
bp	Base pair
BUSCO	Benchmarking Universal Single-Copy Orthologs
CD-HIT	Cluster Database at High Identity with Tolerance
COVID-19	Coronavirus Disease 2019
COX-1	Cyclooxygenase-1
CPR	Central Pharmaceutical Reference Database
CRP	C-Reactive Protein
DDBJ	DNA Data Bank of Japan
DEG	Differentially Expressed Gene
DFD	Data Flow Diagram
DNA	Deoxyribonucleic Acid
EDH	Electronic Health Care Database
eggNOG	Evolutionary Genealogy of Genes
EMBL	The European Molecular Biology Laboratory

EMBL-EBI	The European Bioinformatics Institute
ENA	European Nucleotide Archive
ERD	Entity-Relationship Diagram
ETL	Extract-Transform-Load
FAERS	FDA Adverse Event Reporting System
FDA	U.S. Food and Drug Administration
FFQ	Food Frequency Questionnaire
FRD	False Discovery Rate
FT	Functional Target
FTP	File Transfer Protocol
Gb	Gigabase
GB	Gigabyte
GEO	Gene Expression Omnibus
GePaRD	The German Pharmacoepidemiological Research Database
GO	Gene Ontology
GOS	Global Ocean Sampling
GPU	Graphics Processing Unit
GSEA	Gene Set Enrichment Analysis
HDFS	Hadoop Distributed File System
HDL	High-Density Lipoprotein
HNLC	High-Nitrate Low-Chlorophyll
HPC	High Performance Computing
ICD-10-CM	International Classification of Diseases 10^{th} revision - Clinical Modification
ICD-10-GM	International Classification of Diseases 10^{th} revision - German Modification
ID	Identifier
IDEFICS	Identification and Prevention of Dietary- and Lifestyle-Induced Health Effects in Children and Infants
IL	Interleukin
KEGG	Kyoto Encyclopedia of Genes and Genomes

LASSO	Least Absolute Shrinkage And Selection Operator
LDL	Low-Density Lipoprotein
LFC	Logarithmic Fold-Change
MetaHIT	Metagenomics of the Human Intestinal Tract
MMETSP	Marine Microbial Eukaryote Transcriptome Sequencing Project
mRNA	Messenger RNA
NADPH	Nicotinamide Adenine Dinucleotide Phosphate (reduced)
NCBI	The National Center for Biotechnology Information
NFDI	National Research Data Infrastructure
NGL	Naïve Group Lasso
NOAC	Non-vitamin K (or Novel) Oral Anticoagulant
OGL	Overlapping Group Lasso
OOM	Out Of Memory
OPS	Operationen- und Prozedurenschlüssel (Operation- and Procedure Code)
ORF	Open Reading Frame
PA	Physical Activity
PAM	Partitioning Around Medoids
PASS	Post-Authorization Safety Studies / Post-Marketing Surveillance Studies
Pfam	Protein Families Database
PLS	Partial Least Squares
PV	Pharmacovigilance
PZN	Pharmazentralnummer (Central Pharmaceutical Reference Number)
REST	Representational State Transfer
RNA	Ribonucleic Acid
RNA-Seq	RNA Sequencing
RSEM	RNA-Seq by Expectation Maximization
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SB	Sedentary Behavior

SD	Standard Deviation
SGB	Social Code Book
SHI	Statutory Health Insurance
SNP	Single Nucleotide Polymorphism
SO	Southern Ocean
SOM	Self-Organizing Maps
SRA	NCBI Sequence Read Archive
STEC	Shiga Toxin-Producing *E. coli*
STITCH	Search Tool for Interactions of Chemicals
Tb	Terabase
TB	Terabyte
TK	Techniker Krankenkasse
TNF	Tumor Necrosis Factor
TSA	Transcriptome Shotgun Assembly
TTD	Therapeutic Target Database
UniProt	Universal Protein Knowledgebase
VLDL	Very Low-Density Lipoprotein
WHO	World Health Organization
XML	Extensible Markup Language

Chapter 1

Introduction

1.1 In the light of data

Finding a common term between banking, weather forecast, fitness, cancer, elections, and climate change is not as difficult as it would seem. Data is the currency of the present world; data collected from various sensors are relied upon for decision making and for compiling strategies for crisis management. However, data and knowledge are not equivalents; the "data step" (i.e., acquisition, integration and transformation) is argued to be the most critical and time consuming step in the data mining and consequently knowledge discovery process (van der Putten, 2010). Therefore, data engineering is needed where data exit.

In 1989, knowledge and data engineering have been considered as the studies related to computer-aided information, data and knowledge management (Ramamoorthy and Wah, 1989). Data engineering covers a wide range of applications including social network analysis, web usage mining, business intelligence, financial fraud detection, and precision medicine, addressing the acquisition and transformation of large collections of data to facilitate information extraction.

Biological data represent a large and a challenging sector of data engineering applications, where being "drowned in data and starved for information" (Brown, 2014) could cost as much as the human life itself. Throughout history, "nothing in biology made sense except in the light of data". Data engineering's fundamental understanding of the data life cycle from acquisition, processing and distribution matches the needs of biological research, where data are typically complex and poorly standardized, and where analytics are often challenged by data dimensionality. To comprehensively grasp the biological data life cycle, and properly address the challenges in biological data curation and analytics, it is crucial to understand the biological data ecosystem.

1.1.1 The diverse ecosystem of biological data

The data landscape in biology is rich and diverse in terms of sources, types and fields of application. First, regarding sources, for instance, (bio)medical data can be collected from people through observations (e.g., batch experiments, examinations, interviews) or from sensors (e.g., wearable devices, sequencers, images, and laboratory measurements). Data are also obtained from large repositories and collections such as sequence repositories and administrative databases (e.g., health insurance databases, cancer registries, biobanks). Moreover, data integration is of major interest in biological research, for example, for investigating the interaction between environmental and genetic factors and its effect on disease etiology and prognosis. To achieve that, the records are to be matched, pseudoymized or anonymized and distributed for statistical analysis (El Emam et al., 2009; Chan et al., 2010; Dey et al., 2018). In addition to integration, biomedical data may feed the Internet of Medical Things (a.k.a., Internet of Health Things or Smart Healthcare) to monitor patients' treatment and general health status (Islam et al., 2015; Baker et al., 2017; Dey et al., 2018). Looking at such an ecosystem, we can categorize biological data based on their source into primary data (e.g., observations from experiments and cohort studies) and secondary data (i.e., databases and repositories of routinely collected data).

Second, concerning data types, different types of biological data are becoming readily available at a reduced cost due to the advances in biological data acquisition systems (e.g., high-throughput platforms in genomics, lipidomics). Since the release of the first sequenced genome in 1995 (*Haemophilus influenzae*) (Fleischmann et al., 1995), the number of sequenced genomes has been increasing exponentially attempting to cover the entire Tree of Life on Earth. Nevertheless, the reduced cost of data acquisition, particularly sequencing, is counteracted by the increasing cost of data storage, processing and analysis (i.e., mapping and variant calling), and, most notably, sharing and privacy (Sboner et al., 2011; Stephens et al., 2015). In addition to the most known types of biological data (i.e., sequence- and image-based), biological data types include relational data as in electronic health care databases spanning record linkage systems (e.g., national disease and death registries), electronic medical records and health care claims databases (Pacurariu et al., 2018). As of 2018, in Europe alone, 34 databases of this type exist, covering, as median, 18.5 years of patient time of five million patients (Pacurariu et al., 2018).

High-throughput omics[1] data and electronic health care databases form the majority of the biological data landscape, which in turn renders biological data inherently high-dimensional. Such data are used in a wide range of biological fields of application. In particular, applications in health (e.g., disease epidemiology and personalized medicine) and environment (e.g., ecology and biogeochemistry) are largely based on knowledge extraction from high-dimensional biological data.

[1]The suffix "omics" refers to: "the measurement of the entire complement of a given level of biological molecules and information". For example, the term "genomics" refers to the quantitative study of

In biological sciences, the pathway from data to information is long. The biological data life cycle is as lengthy and costly as the unpacking of genetic information in biological systems, where genetic material is transcribed and/or translated into functional entities (i.e., proteins). Data curation strategies depend on the data source, type and intended analytics. Observation data from sensors or from the field are to be acquired, transferred to storage, normalized (e.g., gene expression data), curated and imputed (e.g., epidemiological surveys), and transformed into features (e.g., gene expression estimates, microbial abundances, drug and disease exposure from electronic health care databases) variables. Such extensive data (pre-)processing requirements are integral in order for the data input type to conform with the intended analytics for knowledge extraction.

Each data source, type and field has its own challenges in data acquisition, management, privacy, processing and analysis. Moreover, with the increasing availability of data sources, data integration is expected, and analysis methods should be scalable to accommodate heterogeneous data sources.

1.1.2 In health and environment

In this subsection, I briefly introduce four challenges in biological data curation and analytics that this thesis addresses, with respect to data source (primary, secondary), and data type (sequence and relational) in the two major fields of application, health and environment.

Knowledge extraction from primary high-dimensional sequence data

Transcriptomic studies of batch culture experiments are a major and popular source of biological primary data in environmental research. They provide a glimpse on the metabolic potential of algae, and a high-resolution snapshot of the metabolism under changing growth conditions. Transcriptomics data acquisition (e.g., RNA-Seq), processing and analysis methods are becoming increasingly standardized (Angiuoli et al., 2008; Osborne et al., 2014; Conesa et al., 2016). Nevertheless, handling high-throughput sequence data represent a challenge to a classical bioinformatic analysis primarily due to dimensionality issues. Moreover, curation and analysis largely depend on the biological source of the data (i.e., its genome complexity and/or genome availability) creating a unique challenge in every study. It is tempting to solve the problem fully and solely automatically, however, subject-matter knowledge and manual curation are often required for knowledge extraction. In spite of being a routine activity with respect to data engineering, applying state-of-the-art

the genome, as in protein-coding genes, regulatory elements and noncoding sequences. Helpful reference: Schneider MV, Orchard S. Omics Technologies, Data and Bioinformatics Principles. In: Mayer B, editor. Bioinformatics for Omics Data: Methods and Protocols. Totowa, NJ (US): Humana Press; 2011. p. 3–30. Available from: https://doi.org/10.1007/978-1-61779-027-0_1.

methods for knowledge extraction from primary sequence data is necessary to address perspective scalability and automation challenges with the increased breadth of available data.

Scalability in handling secondary sequence data

In addition to the various high-resolution "vertical" data types available on an organism, there is the "horizontal" aspect of biological data. Large volumes of data are collected on individual or communities of organisms through global projects. Examples of such projects include: Global Ocean Sampling (GOS) (Yooseph et al., 2007), *Tara* Oceans and Oceanomics (Sunagawa et al., 2015), and the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Keeling et al., 2014). To showcase the data volume by such projects, *Tara* Oceans expeditions yielded 7.2 terabases (Tb) of ocean microbiome data, and a comparable amount of data was yielded by the US Human Microbiome Project and the European Metagenomics of the Human Intestinal Tract (MetaHIT) project (Qin et al., 2010; Human Microbiome Project Consortium, 2012; Li et al., 2014b; Sunagawa et al., 2015). These data are made available through public repositories, which are a valuable secondary data source for information integration and meta-studies. A key challenge in handling secondary sequence data is the scalability and modularity of data acquisition and analysis, to address 1) the differences in sequencing technology and, consequently, in pre-processing steps, and 2) the type of integration and, therefore, appropriate analytics.

Integration and statistical modeling of secondary high-dimensional relational data

It is evident that sequence data analysis (i.e., functional annotation) largely relies on data integration from annotations platforms. Nevertheless, utilizing molecular-based ontologies to analyze relational epidemiological data is foreseen. The recent steep rise in routinely collected health data sources (i.e., electronic health care databases) renders such secondary data a readily available and cheap, yet highly valuable data type for pharmacoepidemiological research, for example, to monitor drug safety in large populations in the post-marketing phase. Handling secondary high-dimensional relational data is challenged at two levels: 1) the extraction and integration of molecular-relevant ontologies from public knowledge bases, and 2) the optimization of statistical methods for scalability, specially for studies on large populations. In addition, data security is a major concern when handling human data.

Integration of heterogeneous primary epidemiological data

Epidemiological research, as well, benefits from the advances in high-throughput multi-omics technologies, which increase the depth of the phenotypic profiles of

individuals, and, in turn, advance our understanding of disease etiology. For optimal knowledge extraction, modern cohort studies require integration of various heterogeneous data types (i.e., lifestyle variables from epidemiological profiling and surveys, exercise data from wearables, and food intake data from food surveys and food tracking web applications), in addition to multi-omics data (e.g., genetic variants, microbiome and lipidome data). Each of these data types are to be transformed prior to integration into a meaningful data model. In the near future, cohort studies will not only be challenged due to the increase in number of variables, but also due to the increase in number of subjects [e.g., as in the UK Biobank including more than 500,000 participants (Sudlow et al., 2015)], calling for scalable solutions for data integration and analytics.

1.2 Thesis objectives, structure and publications

1.2.1 Thesis objectives and structure

The aforementioned data challenges in biological research invite the investigation of scalable workflows to automate acquisition and integration. The role of data engineering in handling high-dimensional data is not limited to data preparation and warehousing, as it extends to closing the gaps in optimizing analytics specially for heterogeneous data.

This thesis, thus, aims at exploring and optimizing the state-of-the-art methods for heterogeneous data integration and analysis, of sequence and non-sequence-based data in human and environmental research. The thesis will present challenges in biological data preparation and transformation for variable selection, where no one-size-fits-all solution can be adopted, and custom-made solutions are required. For this purpose, I identified four areas of application in primary and secondary data concerning a wide spectrum of the Tree of Life (e.g., marine algae, land plants and human gut microbiota). The particular aspects and models for data processing and analysis, and the areas of application are:

1. State-of-the-art dimensionality reduction practices and their impact on knowledge extraction in transcriptomics using a case of primary environment data (Chapter 2)
2. A simple approach for scalability and reproducibility of acquisition and analysis of transcriptomic data from public repositories: Meta-analysis of secondary environment data (Chapter 3)
3. Integration and statistical modeling of high-dimensional relational data for adverse drug event prediction in secondary health care claims data (Chapter 4)

4. A model for meaningful data integration using dimensionality reduction methods: An epidemiological case study of primary heterogeneous multi-omic-based data (Chapter 5)

In these four chapters, I present the four concepts and accompanying case studies. In each chapter, I first present a background on the biological research data sources, acquisition and analytics model. Second, I address the foreseen challenges and requirements, the solution implementation steps, and the evaluation and limitation of the solution. In Chapter 6, I present a conclusion and future outlook. The thesis chapters are complemented with two appendices: a technical appendix A for supplementary information and results, and the publication appendix B for the manuscripts the thesis contributed to, listed below.

1.2.2 List of manuscripts and statement of contribution

1. Manuscript I (Published): Rizkallah MR, Frickenhaus S, Trimborn S, Harms L, Moustafa A, Benes V, Gäbler Schwarz S, Beszteri S. Deciphering patterns of adaptation and acclimation in the transcriptome of *Phaeocystis antarctica* to changing iron conditions. J Phycol 2020; 56: 747–760.

    - I maintained and inoculated the cultures, harvested the cells and extracted RNA with S. Beszteri. I have executed the transcriptome assembly, analysis and differential expression inference with Harms L supervised by S. Frickenhaus and A. Moustafa. I participated in the conceptualization of the manuscript, and I wrote the initial draft of the manuscript with S. Frickenhaus and S. Beszteri. I processed the sequence-based data to be deposited and publicly available through NCBI. The work was done in collaboration with and under the supervision of the co-authors. The manuscript is published in the Journal of Phycology.

2. Manuscript II (Published; in German): Foraita R, Dijkstra L, Falkenberg F, Garling M, Linder R, Pflock R, Rizkallah MR, Schwaninger M, Wright MN, Pigeot I. Detection of drug risks after approval: Methods development for the use of routine statutory health insurance data. Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz 2018; 61: 1075–1081.

    - R. Foraita planned the concept and wrote the main draft of the publication. I and R. Foraita conducted the literature research, prepared visualizations and wrote the initial draft of the section concerning using functional targets-based analysis and construction of patient risk profiles in identifying risk factors for adverse drug events in routine statutory health insurance data.

3. Manuscript III (Draft): Rizkallah MR, Dijkstra L, Wilhelm AFX, Pigeot I, Foraita R. Predicting patient risk for adverse drug events in health care claims data using functional targets.

- R. Foraita, I and L. Dijkstra planned the concept. I conducted the literature research, data curation and transformation pipeline. Statistical analysis plan was compiled by R. Foraita and I in collaboration with L. Dijkstra, I. Pigeot and AFX. Wilhelm. I optimized the methods performance with R. Foraita and L. Dijkstra. I wrote the initial draft, R. Foraita revised parts of the draft.

4. Manuscript IV (In preparation): Wolters M, Rizkallah MR, Foraita R, Liebisch G, Veidebaum T, Tornaritis M, Molnár D, Eiben G, Rampelli S, Günter K, Marron M on behalf of the IDEFICS/I.Family and MyNewGut Consortia. Plasma lipidome and gut microbiome profiles as predictors of weight gain in children.

    - The MyNewGut Consortium planned the concept. M. Wolters, I and R. Foraita conducted the literature research. M. Wolters, K. Günter and I conducted the data selection. I conducted data curation and transformation, and developed the analysis plan with R. Foraita and the MyNewGut Consortium. R. Foraita and I compiled and optimized the statistical methods. I participated in the manuscript writing.

1.2.3 Further contributions

- Aziz RK, Hegazy SM, Yasser R, Rizkallah MR, ElRakaiby MT. Drug pharmacomicrobiomics and toxicomicrobiomics: From scattered reports to systematic studies of drug–microbiome interactions. Expert Opin Drug Metab Toxicol 2018; 14: 1043–1055.
- Aziz RK, Rizkallah MR, Saad R, ElRakaiby MT. Translating pharmacomicrobiomics: Three actionable challenges/prospects in 2020. Omi A J Integr Biol 2020; 24: 60–61.

Chapter 2

Existing Workflows for Knowledge Extraction: A Case of Primary Environment Data

Transcriptomic studies of batch culture experiments are a major source of biological data, and they fall under the umbrella of primary data. These data are, by definition, intended to answer a particular scientific question, such as the adaptation of species under novel growth conditions. As the data acquisition methods (e.g., RNA-Seq) in these experiments are meanwhile highly standardized, curation (i.e., processing) and analysis methods are becoming increasingly standardized as well. Workflows by R-Bioconductor (Love et al., 2015) and best practices (Conesa et al., 2016) for analyzing these data are available. Nevertheless, curation and analysis largely depend on the biological source of the data (i.e., genome complexity and/or availability of the organism, how well-studied the organism is), the biological research question and the skill set of the analysts, creating a unique challenge in every case study. Therefore, different types of downstream analyses or different ontologies might be required. RNA-Seq data are high-dimensional data and thus preparation and analytics are often challenged by dimensionality issues.

In this chapter, I present the analysis of primary data from environmental research, in particular marine algae. The case study is a transcriptomic study of a batch culture of the Southern Ocean key endemic species *Phaeocystis antarctica*. The chapter deals with data acquisition from source, annotation data acquisition, dimensionality reduction, basic data flow management, and study data archiving in public repositories. The work featured in this chapter is published in the *Journal of Phycology* (Appendix B.1).

Although the project seems to be more of a routine activity with respect to data engineering, it paved the way to considering more sophisticated scalable solutions in the following chapters. In particular, due to the popularity of transcriptomics in algal research, in Chapter 3 I apply simple, scalable solutions on secondary sources of transcriptomic data (i.e., sequence public repositories). In addition, the case study

described here offered an opportunity to highlight challenges and solutions for measuring and managing data acquired from a biological experiment.

The chapter is structured as follows. In Section 2.1, I describe the potential of transcriptomics in marine research and on the ecological relevance of the species chosen for this case study before emphasizing the objective of the chapter. In Section 2.2, I address the objective in detail in terms of project requirements, steps undertaken for solution implementation and evaluation of the results. Concluding remarks are presented in the last section.

2.1 Background

2.1.1 Transcriptomics as a potential for marine research

Transcriptomic studies present an interesting source of biological data. These studies can compare gene expression under two or more conditions (e.g., health/disease, enriched/starved) by allowing the organism to grow under different conditions. If applicable, biological replicates of the organism can be used. In case of algae, cells are harvested from replicates and conditions on filters. Then, mRNA[1] is extracted from the cells and sequenced (e.g., by high-throughput technologies such as RNA sequencing; RNA-Seq). In RNA-Seq, total mRNA is sequenced yielding usually short reads (in case of Illumina; 30-150 bp) (Marguerat and Bähler, 2010). To quantify gene expression, these short sequences are mapped to the genome (if available), or *de novo* assembled into a scaffold (as genes). Finally, gene expression values are statistically tested to infer the differentially expressed fraction of the genome under the predefined growth conditions. Many tools and workflows were developed for processing data from high-throughput technologies (Reuter et al., 2015; Conesa et al., 2016), providing guidelines to clean (i.e., process) data resulting from each sequencing technology, to assemble the sequences, and to infer differentially expressed genes and isoforms (i.e., transcripts), in addition to experimental design for plant biology (Strickler et al., 2012) and statistical aspects of RNA-Seq data analysis (Yendrek et al., 2012).

Marine algae represent both a potential and a challenge for data curation and analysis. First, genome organization is generally very complex and genomes are large (e.g., dinoflagellates). Dinoflagellates have some of the largest known genomes of sizes ranging from 1.5 to 185 Gb (gigabases or one billion nucleotides) (Wisecaver and Hackett, 2011). To relate to those numbers, the size of the human genome is

[1] RNA stands for **r**ibo**n**ucleic **a**cid. In the classical view of the central dogma in biology, DNA holds the genetic information. When needed, the information is transcribed into the small portable mRNA (messenger RNA). For more information on the central dogma and how biological data are organized within the databases in relation to the central dogma, I refer to material from NCBI (The National Center for Biotechnology Information). URL: https://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/central_dogma.html. Helpful references: Barrett T, et al. BioProject and BioSample databases at NCBI: Facilitating capture and organization of metadata. Nucleic Acids Res. 2012;40(D1):57-63. Crick F. Central dogma of molecular biology. Nature. 1970;227(5258):561-563.

$\sim$3 billion nucleotides, which translates to $\sim$140 gigabytes of raw data generated from a genome sequencer (Marx, 2013). Large genomes sizes can be a result of gene duplication and lateral gene transfer events [e.g., dinoflagellates (Wisecaver and Hackett, 2011)] or whole-genome duplication [e.g., diatoms (Parks et al., 2018)], which in turn makes genome sequencing and analysis challenging and thus information on genome sequence scarce. Second, unique structural features (e.g., thick silica shells in diatoms) make genetic engineering techniques less utilized in algal molecular biology, which, contributes to the fact that less information is available about the genetics and, consequently, the metabolic potential of these species.

Transcriptomics can serve as an alternative for genome sequencing, as it provides a glimpse on the metabolic potential of algae and a high-resolution snapshot of the metabolism under particular growth conditions. However, there are two issues to account for: 1) the small sample size (sometimes of one) that challenges both the inference of differentially expressed genes and the coverage of sequenced genes, 2) the poor and misleading annotation of assembled genes because the majority of algal species is underrepresented in public databases.

2.1.2 Case study: *Phaeocystis antarctica* and iron metabolism

Iron is essential for phytoplankton growth as it serves as an electron carrier in photosynthesis and mitochondrial respiration. It is also required as a cofactor in fatty acid biosynthesis, nitrate reduction and assimilation (Marchetti et al., 2012; Harel et al., 2014; Schoffman et al., 2016). The effect of iron limitation has only been studied in temperate diatoms at the molecular level (Strzepek and Harrison, 2004; Allen et al., 2008; Lommer et al., 2012). Even though less studied than diatoms, studies on haptophytes demonstrated similar adaptation of haptophytes to iron limitation and lower iron requirements for growth (Strzepek et al., 2011, 2012).

Phaeocystis is a cosmopolitan genus within the division of haptophytes. Its most famous members are three colony- and bloom-forming species: the temperate *P. globosa* in the North Sea, the Arctic *P. poucheti* and the Antarctic *P. antarctica* (Schoemann et al., 2005; Verity et al., 2007). Colonial life stage provides those species with protective and competitive advantages over the solitary stage, with the protein-carbohydrate colony skin serving as a mechanical barrier against infections, and the large colony size protecting against grazers. The mucilaginous structure of the colonies matrix further allows for storage of micro- (iron and manganese) and macro- (carbon and nitrogen) nutrients (Hamm, 2000; Schoemann et al., 2005; Gaebler-Schwarz et al., 2010).

The Southern Ocean (SO) is the largest high-nitrate low-chlorophyll (HNLC) region with subnanomolar concentrations of total dissolved iron and abundant concentrations of macronutrients yet low rates of nitrate uptake, and dominance of pico- and nanophytoplankton species (Dugdale and Wilkerson, 1991; Smetacek et al., 1997;

Assmy et al., 2007). Iron supply to the SO includes dust deposition and melting icebergs (Assmy et al., 2007), but as iron remains bound to organic ligands and therefore biologically unavailable to phytoplankton (Shaked and Lis, 2012; Hutchins and Boyd, 2016), it is limiting phytoplankton growth and productivity (Martin et al., 1990).

Phaeocystis antarctica is endemic to the largely iron-limited SO and forms large *P. antarctica* blooms, which are frequently recorded in the iron-enriched shelf areas such as Ross Sea and Prydz Bay (Boyd, 2002a; Schoemann et al., 2005; Smith et al., 2014b). *In vitro* experiments showed that *P. antarctica* has a strong response to iron limitation as indicated by reduction in its growth rates and photosynthetic fitness (Strzepek et al., 2011; Alderkamp et al., 2012), while iron addition was reported to increase growth rates, and trigger colony formation in *P. antarctica* (Bender et al., 2018). *In situ* iron fertilization experiments in the SO reported haptophytes (*P. antarctica*) among the groups contributing to the detected peak of photosynthetic activity (measured as the elevation in chlorophyll a signal) after iron enrichment (Gall et al., 2001; Boyd, 2002b; de Baar et al., 2005). In the subarctic Pacific, metatranscriptomics showed that haptophytes (*P. globosa*) utilized added iron faster than diatoms, with an overexpression of photosynthesis genes (Marchetti et al., 2012).

2.1.3 Study objectives

The **biological objective** of this case study was to decipher the molecular adaptation to low iron availability and subsequent acclimation following iron enrichment in the ecologically important prymnesiophyte *Phaeocystis antarctica*. This is achieved by conducting a transcriptomic study on a colony-forming isolate from the Ross Sea. The results of this study highlight the molecular processes that might be the basis of the adaptation of *P. antarctica* to iron limitation, and its acclimation to iron addition. For a detailed description of the biological results, refer to Appendix B.1.

I use this typical transcriptomic study to illustrate the process of data dimensionality reduction as one important area of **data engineering**. For this purpose, I employed dimensionality reduction methods throughout the steps of data pre-processing, clustering and differential expression analysis and evaluated their effect on knowledge extraction. In this chapter, I highlight the importance of both automation of data preparation and domain knowledge in knowledge extraction from primary data. I also briefly describe the workflow of the software I utilized, Trinity (Grabherr et al., 2011; Haas et al., 2013), according to the **E**xtract-**T**ransform-**L**oad (ETL) workflow criteria, and pinpoint areas for robustness and scalability.

2.2 Dimensionality reduction

As mentioned above, data dimensionality is a challenging aspect of managing and analyzing biological data. In this section, I focus on dimensionality reduction discussing: 1) the data engineering and statistical challenges in the light of the project requirements, 2) the steps undertaken for solution implementation, 3) the results and evaluation of the data preparation and analysis solutions, and 4) limitations of this case study, and of my solution and evaluation. I also briefly describe the main software used in this study from an ETL perspective.

2.2.1 Challenges and project requirements

There is a set of core requirements for data preparation (i.e., pre-processing) and analysis in a transcriptomic study, in addition to study-specific challenges. A set of core requirements for a transcriptomic study are discussed in detail in (Conesa et al., 2016). These requirements cover aspects of pre-processing (experimental design, sequencing and quality control), core analysis (transcriptome profiling and differential expression) and advanced analysis (visualization and integration of other omics data types).

In this case study, it was important to not only quantify the expression at different growth conditions, but also to construct a draft transcriptome to demonstrate the metabolic potential of the species. Here I summarize the requirements for this particular case study following the order of the data flow in transcriptomics.

1. **Sequence pre-processing.** RNA sequences cannot be used unless cleaned from (i.e., trimmed of) sequencing adapters and low-quality bases due to sequencing errors. Depending on the sequencing technology, the respective software and adapters must be used, see Chapter 3 for more information on different technologies and software for reads pre-processing. In the case study, extracted RNA was sent for paired-end RNA sequencing using Illumina HiSeq2000 sequencer. Therefore, Illumina-specific adapters and software were sought.

2. **Transcriptome profiling: *De novo* assembly and annotation.** The short RNA-Seq reads have to be transformed into "full-length" transcripts. This can be achieved by either mapping the reads onto the genome through alignment [evaluated in (Engström et al., 2013)], or assembly of the transcripts as in Trinity (Grabherr et al., 2011; Haas et al., 2013). As the genome of *P. antarctica* is not yet available, a *de novo* assembly of the reads is required to transform the RNA-Seq reads into transcripts. A transcriptome assembler has to be able to resolve not only the expressed genes but also alternative splicing events if any (Grabherr et al., 2011). In addition, to increase the depth of the assembly, all sequenced replicates are better assembled together. Assembled transcripts have to be assigned a function (i.e., annotated), generally based on sequence similarity to

known genes in public data repositories (e.g., UniProt[2]), and based on domain search (e.g., Pfam[3]). The best annotation then has to be chosen based on, for instance, the statistical significance of the match between the query and the similar sequences in the database (a.k.a., "the hits"). In addition, while conducting the study, the transcriptome of *P. antarctica* from The Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP) was released (Koid et al., 2014). As I wanted to compare the study transcriptome to that from MMETSP, a metric had to be employed.

3. **Expression quantification.** In biology, the purpose of transcriptomic studies is to quantify and compare the gene expression under tested conditions (e.g., health and diseased; treated and untreated). Therefore, a main step of transcriptomic studies is the estimation of gene and/or transcript/isoform expression levels under each condition. These data are often deposited in public repositories (e.g., Gene Expression Omnibus database; GEO). The abundance can be estimated for transcripts, genes and most recently "supertranscripts" [i.e., all the exons of a gene (Davidson and Oshlack, 2018)]. Expression quantification can currently be achieved through alignment-based or alignment-free methods[4]. In this case study, *P. antarctica*, like many other non-model organisms, has no published genome, therefore, a method for gene and isoform quantification from only RNA-Seq data without the need of a reference genome had to be utilized.

4. **Differential expression and functional profiling** of statistically significantly different transcripts between the samples is the final step of a transcriptomic study. There are two issues here: 1) the selection of the information level for differential expression (i.e., genes or transcripts; observations), and 2) the selection of the conditions that need to be compared (i.e., samples). In this case study, RNA was collected from the inoculum, quadruplicates before and after iron addition (morning and evening), the informative samples (time points) are those that are available in replicates (thus exclude inoculum) and available at comparable times (all evening). Furthermore, dimensionality reduction is required, as in inferring patterns of expression (e.g., using k-mean clustering), and to characterize the molecular pathways in which the differentially expressed genes are involved (e.g., using gene-set enrichment analysis). Statistical significance does not imply biological relevance per se; domain knowledge is essential to distinguish between the observations that are relevant and those that are merely significant in the statistical sense.

[2]Bateman A, et al. UniProt: The universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):D158–69. URL: https://www.uniprot.org/

[3]El-Gebali S, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47(D1):D427–32. URL: http://pfam.xfam.org/

[4]Haas B. Trinity GitHub repository. URL: https://github.com/trinityrnaseq/

Existing solutions

To fulfill the aforementioned requirements, commercial state-of-the-art solutions are sought. In particular, at the time of the conception of this study, Trinity (Grabherr et al., 2011) was recently released and evaluated as a suitable *de novo* assembler for non-model organisms [e.g., algae (Cohen et al., 2018; Koch et al., 2019)]. Arguments for choosing Trinity include: 1) ease of setup, use and interpretation as it is extensively documented, and 2) highest assembly quality scores and highest read alignment rates (Wang and Gribskov, 2017). Moreover, alongside Trinity assembler, software solutions for data pre-processing, data flow management and downstream analysis were developed and became easy to use in conjugation with Trinity. Trinity ad-hoc tools include software for: Illumina sequence pre-processing [Trimmomatic (Bolger et al., 2014)], open reading frame (ORF) extraction (TransDecoder) and annotation (Trinotate). In addition, independent methods for expression quantification such as "RNA-Seq by Expectation Maximization" (RSEM) (Li and Dewey, 2011) and for differential expression analysis [e.g., DESeq2 (Love et al., 2014)] are easy to use with Trinity. Taken together, Trinity provides a semi-automated, full-suite solution for *de novo* assembly and analysis of RNA-Seq data. To understand the data structure within a transcriptomic study, in Figure 2.1, I provide a simple illustration showing the data at different stages of the study, namely sequence pre-processing, assembly, annotation, and expression quantification.

Trinity-based solution, however, is challenged by data dimensionality in two ways. Typically, Trinity assembler produces a large number of transcripts (i.e., isoforms of hypothetical genes) with no consensus sequence of the hypothetical genes. Therefore, if the growth conditions are largely different, DESeq2 produces a large number of statistically significant differentially expressed isoforms. Both make it difficult to communicate the high-dimensional data produced by the annotation pipeline Trinotate to analysts and scientists, as Trinotate attempts to annotate each isoform of each hypothetical gene. Therefore, a transcriptomic study using Trinity requires the development of a multi-step dimensionality reduction protocol using Trinity adjustments, and programming-based and statistical methods.

2.2.2 Solution implementation

In this case study, I considered many stages of dimensionality reduction to overcome the large number of transcripts produced by Trinity. Here I describe the levels where dimensionality reductions were employed with respect to: 1) the transcriptomics data flow (Figure 2.2), and 2) the tools used for data preparation and analysis. The diagram illustrates the workflow based on Trinity assembler and its ad-hoc tools using RNA-Seq reads as the input.

1. **Sequence pre-processing.** To trim sequencing adapters and to eliminate low-quality bases and very short reads (Ungaro et al., 2017) (i.e., quality control),

Trimmomatic (v0.32) was used.

2. **Transcriptome profiling: *De novo* assembly and annotation.** To increase the depth of the assembly and construct a comprehensive transcriptome, all sequenced reads were pooled into a single input file for one assembly run by Trinity (v2.0.4). In this file, only quality-filtered paired-end reads were retained. The samples sequenced and used for assembly were: 1) the iron-limited culture that was used to inoculate quadruplicates, 2) the iron-deplete control (quadruplicates pre-iron treatment), and 3) iron-replete treatment (quadruplicates post-iron treatment). These samples were taken at one time point for the controls, and three time points after the treatment [14h ($n = 3$), 24h and 72h]. As for **coding sequence prediction**, ORFs were extracted from the assembled transcripts using TransDecoder (v2.0.1). Only long ORFs (longer than 100 bases) were retained. Moreover, predicted ORFs were screened for homology to known proteins in databases (UniProt and Pfam). Only the ORFs with similarity to known proteins were retained. Regarding **functional annotation**, the translated ORFs (i.e., predicted proteins) were analyzed using Trinotate (v2.0). Only transcripts with known functions (compared against UniProt or Pfam) were considered and mammalian hits were excluded. In case of comparison against UniProt using BLAST, significance of the sequence similarity was inferred by comparing the observed similarity (bit) score with the expected number of sequences in the database that have a bit score at least equal to the observed; called expected (E) value[5]. We chose E-value ≤ 0.00001 as a threshold to be not too permissive, yet allow for discovery of genes of novel functions. Functional annotation was collapsed per hypothetical gene based on best UniProt and Pfam hits of the longest ORF using `Bash`, `R` and `SQL` scripting languages [6]. In addition to standard functional annotation, I conducted a **comparison against published *P. antarctica* transcriptome** using state-of-the-art tools. Three metrics for three criteria were employed: 1) functional coverage (using BUSCO[7]), 2) sequence coverage of published organellar genomes (using BLAST) and 3) sequence overlap (using OrthoMCL[8]). Details on the methods and results are available in the respective sections in Appendix B.1.

3. **Expression quantification** was conducted using RSEM at the gene-level to reduce the number of observations in the downstream analyses.

[5] More details can be found in Appendix A.1.

[6] A custom script. Code is available upon request.

[7] Benchmarking Universal Single-Copy Orthologs (BUSCO) v1.22. Simão FA, et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2. URL: https://busco.ezlab.org/

[8] Ortholog Groups of Protein Sequences (OrthoMCL) v2.0.9. Li L, et al. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89. URL: https://orthomcl.org/orthomcl/

4. **Differential expression and functional profiling of significant genes.** To minimize the number of observations considered for differential expression analysis, we only considered: 1) expression at gene-level and 2) genes of ≥ 300 bases and sum of rounded counts ≥ 40. We then applied cutoffs for: 1) false discovery rate (FDR) ≤ 0.001 and 2) absolute logarithmic fold-change (LFC) ≥ 2 (i.e., the magnitude by which the expression of a gene is affected by the treatment). Regarding the samples, control and treatment replicates were used excluding: 1) the inoculum and 2) the replicate that was considered an outlier according to principal component analysis of normalized expression values. Moreover, for comparability purposes, it was necessary to exclude the genes that were significantly expressed exclusively at the morning time point from all other time points. Significantly expressed genes were further analyzed: 1) to infer patterns of expression across time, k-mean clustering was used, and 2) to understand the biological functions overrepresented in each cluster/pattern, pathway analysis was conducted using gene ontologies eggNOG[9] and GO[10].

2.2.3 Evaluation

In transcriptomics, primary data are analyzed to answer a particular research question, therefore efficient knowledge extraction from high-dimensional data becomes crucial. In this case study, various dimensionality reduction steps were applied and their outcomes were cumulatively evaluated on the basis of extracted knowledge.

The evaluation criteria include: 1) the quality of the assembly and the comparison of the assembled transcriptome and published *P. antarctica* sequences, 2) the knowledge extracted on differentially expressed pathways (based on eggNOG and GO) and relevant genes (based on domain knowledge) under changing iron conditions.

Assembly

First, for evaluating the quality of the transcriptome assembly, I used the percentage of sequences rejected by NCBI quality control checks. The assembled transcripts were cleaned and submitted to the Transcriptome Shotgun Assembly (TSA) repository (DDBJ/EMBL/GenBank; Accession: GFUQ00000000). As few as 174 transcripts (0.0014%) were rejected. Other data generated from the study were also deposited at respective repositories: 1) metadata (BioProject; PRJNA395466), 2) quality-filtered raw sequencing reads [Sequence Read Archive (SRA); SRP113407] and 3) count and normalized gene expression matrices [Gene Expression Omnibus (GEO); GSE102608].

[9]Evolutionary Genealogy of Genes (eggNOG) v4.0. Powell S, et al. eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Res. 2012;40(Database issue):D284–9.

[10]Gene Ontology (GO). Ashburner M, et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.

Making the project data available was not only important for communicating the results to the scientific community, but it was also an opportunity to interact with a number of public repositories for transcriptomics data.

Moreover, we compared the assembled transcriptome to published *P. antarctica* sequences. 1) We compared the study transcriptome against the published *P. antarctica* transcriptome from the MMETSP project (Koid et al., 2014) (MMETSP1100 containing 53,204 coding sequences and 54,300 peptide sequences) in terms of sequence overlap using OrthoMCL. A relatively small portion of the translated ORFs (9% of whole transcriptome; de-duplicated) were orthologs of 25% of MMETSP's *P. antarctica* coding sequences. The low similarity between the sequences of both transcriptomes can be attributed to the software used (OrthoMCL; it compares the translated predicted coding sequences), differences in sequencing depth in both strains, or to the differences in data processing between the two studies. 2) We compared sequence coverage in both transcriptomes of published partial mitochondrial and complete plastid genomes (Smith et al., 2014a). The study's transcriptome showed better coverage of both plastid (51 non-overlapping transcripts; 93.4% of organellar genome length) and mitochondrial genomes (17; 73.7%), compared to 1% and 0%, respectively, in the MMETSP transcriptome, showing that this assembly worked better than that of MMETSP.

Dimensionality reduction

Second, for evaluating the dimensionality reduction approaches employed, I show the reduction in number of observations, and highlight the extracted knowledge on differentially expressed pathways and genes. The impact of applying dimensionality reduction techniques on the data dimensions illustrated in Figure 2.1 is manifested in Figure 2.3, which depicts the number of observations at each of the dimensionality reduction steps. Quality filtration of the reads led to improving the quality of assembled transcripts, which consequently led to enhancing our knowledge on the metabolic potential of *P. antarctica*. In addition, three components helped the analysts with knowledge extraction the most: 1) using information at gene-level, 2) clustering of differentially expressed genes, and 3) using pathway analysis (eggNOG); details in Appendix B.1. In addition, a crucial step in the study was employing domain knowledge to mine the genes that are biologically relevant to the ecophysiology of *P. antarctica*. Therefore, standard and simple dimensionality reduction techniques, effective visualization of clustered genes and knowledge-based assessment of the differentially expressed genes had a great effect on knowledge extraction.

The knowledge on differentially expressed pathways and genes in *P. antarctica* under iron-limited and -replete conditions were linked to physiological observations of *P. antarctica* of the batch culture and supported by reported *in situ* and *in vitro* observations. First, we observed an up-regulation in genes involved in photosynthetic activity, mucus formation and down-regulation of motility and motor/flagellar genes.

These observations match the increase in photosynthetic fitness and abundance of colonial cells observed in the cultures after iron addition (Issak, 2014). Similarly as in the literature, colony formation and blooms were recorded in the iron-enriched shelf areas (Boyd, 2002a; Schoemann et al., 2005; Smith et al., 2014b) and after iron addition (Bender et al., 2018). Likewise, photosynthetic pigment production was elevated after iron enrichment (Gall et al., 2001; Boyd, 2002b; de Baar et al., 2005). Methods such as RT-PCR are required to validate the expression of marker genes under changing iron conditions. Our results also suggested three adaptive strategies that *P. antarctica* may have utilized under low iron availability: 1) activation of an alternative growth mode (mixotrophy) as per the observed overexpression of motility and endocytosis genes under low iron, 2) expression of iron-economic alternatives of key enzymes (e.g., flavoxdoxin instead of ferredoxin for photosynthesis), and 3) expression of iron-independent functional alternatives of other crucial enzymes (e.g., NADPH-dependent nitrite reductases for nitrite metabolism). Nevertheless, more investigation is required to test our hypotheses, which was beyond the scope of this exploratory project.

Viewing Trinity from an ETL perspective

Trinity presents a semi-automated, full-suite open-source and free solution for RNA-Seq analysis. Trinity and its ad-hoc tools can be considered as a decision support system (Henry et al., 2005), as they manage data flow, analysis, visualization and integration (i.e., measurement and management) of RNA-Seq data. In particular, the annotation tool Trinotate integrates data from different annotations platforms into a central relational database and transforms it into a tabular format for analysts. The differential expression workflow of Trinity transforms data and reports results using powerful visualizations. In addition, Trinity handles most of the parallelization required for working with high-throughput data.

Given their importance and rise, Henry *et al.* (Henry et al., 2005) have developed a trade study for ETL tools evaluation. This thorough study contains criteria, figures of merit, test scenarios and quantitative measures for evaluation. I used the figures of merit for measuring the quality of Trinity from an ETL perspective. According to these figures of merit, Trinity's strengths are: speed, flexibility, cost and ease of use. Trinity is, however, challenged in the areas of robustness and scalability.

First, all hardware and software requirements of Trinity are listed and documented. Nevertheless, software dependencies of Trinity and its ad-hoc tools such as annotation databases and `R` packages are all required to be independently installed *apriori*. Such issue hinders synchronization and dynamic extraction of data, which can lead to, for example, outdated functional annotation due to the use of older versions of annotation databases. Second, Trinity can handle projects of varying sizes, yet not sequentially; to carry out multiple projects or runs of a single project, automation by

programmers is required. Therefore, there is a potential for improving the robustness and scalability of Trinity perhaps through automation at two levels: 1) dynamic acquisition of annotation databases and update of the results in a reproducible manner, and 2) automation of the required pipeline steps using simple approaches such as scripting languages.

2.2.4 Critical appraisal

Trinity is well-established as an Illumina RNA-Seq data assembler and a solution for RNA-Seq data analysis specially for non-model organisms. Such full-suite solutions give fast results as they smoothly handle measurement and management of RNA-Seq data facilitating results interpretation by the scientific community. Nevertheless, several factors contribute to the high-dimensional nature of the data generated by RNA experiments and Trinity itself.

In this case study, I used a mixture of statistical methods and programming tools to reduce the data dimensions. Statistical methods had more advantages compared to programming-based tools. Statistical methods were easier to explain and communicate, document and verify. It was also easier to report on their parameters and to evaluate their performance. Development of the programming-based methods, however, requires extensive documentation that a skilled engineer would be able to provide.

In this case study, I based my evaluation on quality of the assembly and knowledge extraction, which falls short in quantifying the amount of information lost (or gained) by employing dimensionality reduction approaches. In order to support decision on the appropriate dimensionality reduction approaches, the developed programming tools need to be documented and evaluated at information level (rather than at knowledge level).

The evaluation did not cover an experiment-specific factor that might have contributed to the data dimensionality issue, namely polyploidy in *P. antarctica*. Even though little is known about the morphological features of *P. antarctica*, it has been reported that colonial cells are diploid, while solitary flagellates are either haploid or diploid [investigated in (Gaebler-Schwarz et al., 2010)]. After iron addition, microscopical examination showed a mixed prevalence of solitary and colonial cells, which made it difficult to handle polyploidy. Moreover, the test strain in the case study was isolated in 1992, and it is likely that SNPs have accumulated over time which led to divergent sequences of the same gene. We think that polyploidy and/or mutations played a role in inflating the number of isoforms per gene in our experiment. Among others, programs that produce non-redundant representative sequences (e.g., CD-HIT[11]), would be employed and evaluated to assess the impact of such phenomena on data dimensions.

[11] CD-HIT: Cluster Database at High Identity with Tolerance. URL: http://www.bioinformatics.org/cd-hit/

2.3 Concluding remarks

Transcriptomic studies using batch cultures are a rich source for primary biological data, which are typically high-dimensional. Transcriptomic studies of non-model organisms represent an opportunity for advancing environmental research. For instance, studies on marine algae help understanding the metabolic potential of these major players in global carbon and sulfur cycles and consequently their effects on the global climate and food chain.

Even though methods of data acquisition and curation are standardized, this case study shows that there is a number of aspects that might aggravate data dimensionality issues and hinder RNA-Seq data preparation and analysis. These aspects include: challenged functional gene annotation, limited information on alternative splicing in the target organism, large differences in growth conditions, the project requirements and software used, in addition to the skill set of the analysts. This case study also illustrates that existing workflows (Trinity) provide fast, easy-to-interpret results as they smoothly handle measurement and management of RNA-Seq data. However, it is necessary to employ programming and statistical dimensionality reduction methods to optimize knowledge extraction. Moreover, evaluation of the dimensionality reduction approaches must address the information loss and/or gain.

Due to both the importance of transcriptomic studies and the advances in RNA-Seq technologies, a large chunk of the information available on marine algae in public data repositories comes from RNA-Seq data. Moreover, as this case study shows, however functioning, existing workflows and employed standard dimensionality reduction approaches are challenged in areas of effective documentation, robustness and scalability. In Chapter 3, I develop a simple, scalable solution for analyzing data from a number of marine transcriptomic studies acquired from public data repositories. This would facilitate expanding the view I acquired from this case study on both improving the usage scalability of existing RNA-Seq suite-solutions and comparing the effects of iron limitation among well-studied organisms (land plants) and their evolutionary ancestors and closely-related organisms (algae).

A) Example of sequenced reads and quality data (FASTQ)

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""""7F@71,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==
```

@title and optional description
sequence line(s)
+optional repeat of title line
quality line(s)

B) Example of reads assembled into transcripts (FASTA)

```
>TRINITY_DN1000_c115_g5_i1 len=247
AATCTTTTTTGGTATTGGCAGTACTGTGCTCTGGGTAGC
TGATTAGGGCAAAAGAAGACACACAATAAAGAACCAGGT
GTTAGACGTCAGCAAGTCAAGGCCTTGGTTCTCAGCAGA
```

C) Headers of the tabular output of the annotation pipeline (Trinotate)

```
0       #gene_id
1       transcript_id
2       sprot_Top_BLASTX_hit
3       RNAMMER
4       prot_id
5       prot_coords
6       sprot_Top_BLASTP_hit
7       custom_pombe_pep_BLASTX
8       custom_pombe_pep_BLASTP
9       Pfam
10      SignalP
11      TmHMM
12      eggnog
13      Kegg
14      gene_ontology_blast
15      gene_ontology_pfam
16      transcript
17      peptide
```

D) Representation of a two-dimensional array containing gene expression data (Count data)

	sample_1	sample_2	sample_3
gene_1	29.00	44.00	16.00
gene_2	0.00	0.00	3.00
gene_3	7.00	6.00	6.00
gene_4	369.00	95.00	135.00

FIGURE 2.1: Data structure representation of different stages of a transcriptomic study. **A** is from (Cock et al., 2009); it represents sequence data in FASTQ format. **B** is from Trinity documentation (URL: https://github.com/trinityrnaseq/); it represents an example of Trinity's output of assembled reads. **C** is from Trinotate documentation (URL: https://github.com/Trinotate/); it represents the fields contained in the annotation report. **D** is a simplified matrix of expression quantification data. The matrix dimensions of **D** in this case study are: 110,971 hypothetical genes x 16 sequenced samples.

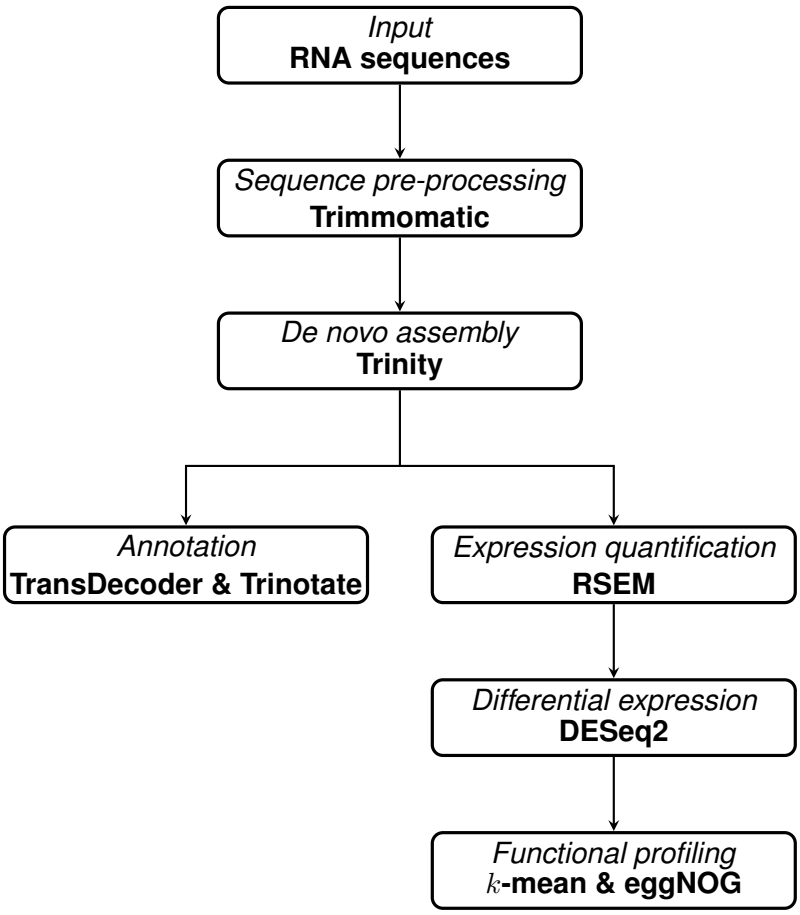


FIGURE 2.2: A simplified illustration of data pre-processing and analysis processes undertaken in this study. In each box, the process (upper) and the tools (lower) are stated.

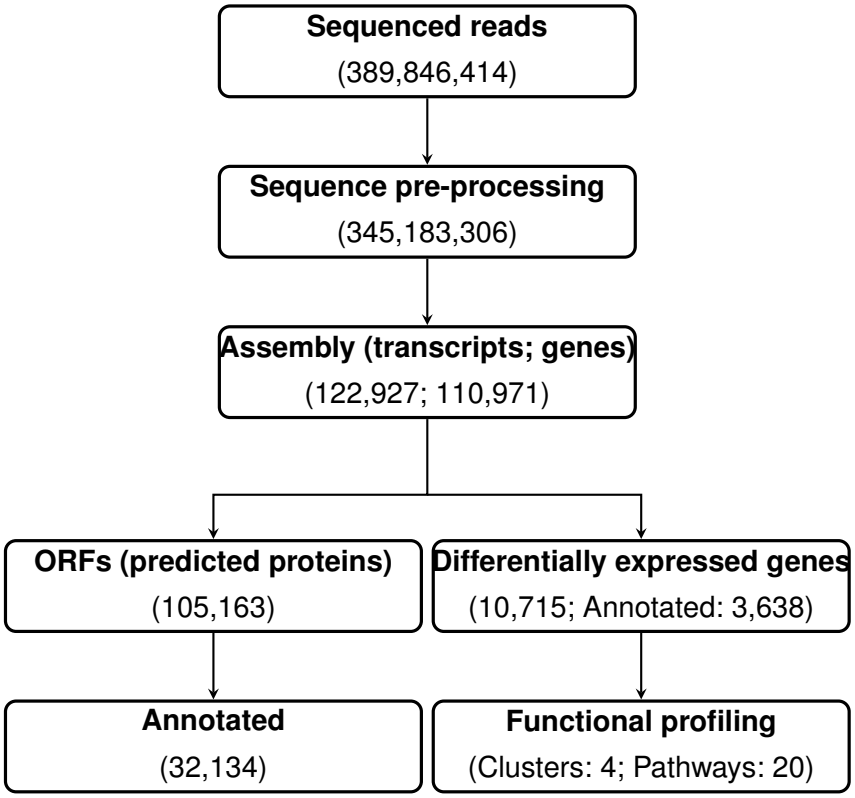


FIGURE 2.3: An illustration of the number of observations that each dimensionality reduction step undertaken in this chapter resulted in. Note that differentially expressed gene statistics are provided at gene-level.

Chapter 3

Scalability and Information Integration: A Meta-Analysis of Secondary Environment Data

Gene expression profiling is one of the most important tools in understanding the metabolic state of the cell. The advances in high-throughput sequencing technologies, and the consequent rise in availability of gene expression data on non-model organisms, could expand our knowledge of the biology of integral members of the Tree of Life. Those members are key players in the global biogeochemical cycles and their response to changing environmental conditions. Potentially, information integration of the less studied non-model organisms with the more studied closely-related organisms (e.g., cyanobacteria and land plants) can sharpen our view on essential biological processes under environmental stressors. Public repositories of sequencing data present a valuable source of secondary data on those organisms.

Data from the two most popular high-throughput sequencing technologies motivate distinct integration schemes. On the one hand, data from RNA-Seq experiments allow for hypothesis generation and exploration of novel genes that are expressed under various environmental stressors. Large high-quality databanks of RNA-Seq experiments on marine algae are becoming available (e.g., Marine Microbial Eukaryote Transcriptome Sequencing Project; MMETSP), which facilitates and motivates information integration. On the other hand, microarray sequencing and analytics protocols are largely standardized, and perhaps better motivate information integration of studies of stressors on the same organism.

Although public repositories ensure accessibility to high-quality validated data, handling secondary molecular data faces challenges on the levels of data acquisition (i.e., transfer, storage and harmonization) and analytics. Moreover, handling these secondary data is largely driven by the gene expression technology used. Therefore, the development of effective data management and suitable analytical approaches are essential to fully decipher the biological knowledge contained in the increasing amount of available sequence data.

This chapter aims at developing a scalable solution for gene expression data acquisition from public repositories, as well as analytics using a case study from environmental research as an example of secondary, routinely collected data. In the case study, I explore the main public repositories for gene expression studies (Gene Expression Omnibus database; GEO) and sequence read data (NCBI Sequence Read Archive; SRA) to acquire, integrate and analyze data from multiple studies on model and non-model organisms using `Bash` and `R` languages. In particular, the case study aims at: 1) comprehensively investigating the change in molecular response in a wide range of photosynthetic organisms (namely: diatoms, haptophytes, green algae, cyanobacteria, land plants) in response to environmental changes in iron availability, and ultimately 2) inferring a core response to iron limitation in photosynthetic organisms. The case study was conceptualized in collaboration with Ahmed Moustafa, Professor of Bioinformatics at the American University in Cairo, Egypt. It will be expanded with evolutionary analysis and prepared for publishing. This work received no funding.

The chapter is structured as follows. Section 3.1 gives background information on gene expression data sources and repositories, dimensions and domains of scalability, and the potential of information integration in studying environmental changes giving the example of the case study. Section 3.2 addresses the chapter objectives with respect to project requirements, solution implementation and results evaluation. Concluding remarks are presented in the last section.

3.1 Background

3.1.1 Gene expression profiling technologies: Data repositories and applications

The metabolic state of the cell is encoded in its transcriptome. Transcriptomic studies, at their core, are set to provide a high-resolution snapshot of that metabolic state, the metabolic potential of the organism under study (see Chapter 2) and to infer reliable biomarkers (Walsh et al., 2015). Cell transcriptome is a very sensitive proxy for the amount of change in environmental conditions affecting the cell (Ogata et al., 2015). This is possibly due to the limitations imposed on gene expression and energy constraints caused by environmental stressors (Wagner, 2005), and due to the limited capacity of the cell that it can respond to a limited number of stimuli (Rhee et al., 2012).

Two main technologies have been used for gene expression profiling: microarray and RNA-Seq [reviewed and compared here (Marguerat and Bähler, 2010; Mantione et al., 2014)]. First, microarray allows for quantifying the expression of annotated and known genes in the cell in, initially, a cost-effective simultaneous manner. Second, RNA-Seq allows for both quantification and detection of novel expressed genes as

well as alternative splicing events at a much higher resolution than microarray. Although throughout the past decade RNA-Seq has become increasingly affordable (Marguerat and Bähler, 2010; Mantione et al., 2014), the cost of data storage and sharing increases dramatically in case of RNA-Seq compared to microarray due to the large volume of raw sequence files produced. Processing and analysis protocols of microarray data are thought to be more standardized than those of RNA-Seq data. Nevertheless, data from a microarray experiment are intended at answering a particular research question, unlike RNA-Seq experiments, whose data can be used (and re-used) to investigate different aspects of gene expression (Mantione et al., 2014).

Regardless of the technology or the study purpose, transcriptomic experiments generally yield a main data type, namely processed normalized transcripts abundance estimates in table format. The data can be deposited in a number of public repositories, reviewed in (Rung and Brazma, 2013), the most well-known of which is the Gene Expression Omnibus (GEO) repository (Barrett et al., 2013). GEO is a public archive of functional genomics data (i.e., raw, processed and metadata). With respect to raw data, RNA-Seq yields raw sequence reads with quality scores, while microarray yields files for scan quantification or intensity calculations of pixel values; both are supported by GEO. Moreover, GEO's functionalities allow the users to query, analyze and download repository data (Barrett et al., 2013). A number of methods were developed for curation and mining of gene expression profiling data archived in GEO [reviewed in (Wang et al., 2019)].

Transcriptomics applications, particularly those utilizing RNA-Seq technologies, extend beyond transcript abundance estimation to investigate transcriptional (e.g., by sequencing short regulatory RNAs) and post-transcriptional (e.g., by analyzing transcript structure and sequence re-arrangement and/or fusion) regulation mechanisms (Marguerat and Bähler, 2010). RNA-Seq captures a large number of expressed genes, which might raise a question whether the gene response we observe is the behavior of one program adopted by each cell in the community or it is a community response [reviewed in (Marguerat and Bähler, 2010)]. Moreover, transcriptomics motivated whole transcriptome-based analyses in uni- and multicellular eukaryotes. These analyses manifested in methodological applications such as: 1) estimation of information content as a function of sequencing depth utilizing Shannon entropy (Kliebenstein, 2012), 2) quantification of transcriptome diversity and specificity also utilizing Shannon entropy (Martínez and Humberto Reyes-Valdé, 2008; Zambelli et al., 2018), and 3) estimation of relative expression per cell, which is particularly important in polyploid organisms (Coate and Doyle, 2010). Transcriptomics were also demonstrated as forensic tool for prediction of time of death based on the expression of marker genes (Hunter et al., 2017). However, being a measurement of abundance and activity of multiple cells in a community, further technologies, such as single-cell ribotyping, were used to correct for expression rate per cell based on

ribosomal RNA copy number, and test for correlation with changes in body size and growth rate in marine protists (Fu and Gong, 2017). Copy number variation has important implications in understanding ecological diversity with respect to biomass rather than cell abundance in response to changing environmental conditions (Fu and Gong, 2017).

3.1.2 Information integration potential and requirements

The increasing availability of transcriptomic data in public repositories constantly motivated information integration from multiple transcriptomic studies, promising improved biomarkers selection, which is of utmost importance in disease and therapeutics monitoring (Walsh et al., 2015). Such improvement is attributed to the increase in statistical power due to the increase in sample size. Levels of public data re-use are perfectly surveyed and described in (Rung and Brazma, 2013). Those levels are: 1) analysis of raw data, 2) meta-analysis of summary-level data (i.e., resulting p-value, effect size or gene rank), 3) supportive analysis combining newly generated and archived data, and 4) performance evaluation of new analytical methods (Rung and Brazma, 2013). Here I focus on the first level.

Integration approaches for multiple raw gene expression data sets are categorized into early- and late-stage integration (Walsh et al., 2015; Frolova and Obolenska, 2016). In early-stage integration (i.e., cross-platform merging and normalization), data from each study are pre-processed, and a unified case-cohort data set is analyzed to identify signature genes. In late-stage integration (i.e., meta-analysis), each case-cohort microarray data set is pre-processed and signature genes are identified and statistically combined. Possible aspects that can impact integration studies include: 1) the research question and whether the platforms are similar, which drive the choice of the cross-platform normalization and the meta-analysis methods, and 2) the transcriptomic data quality, which requires careful pre-processing and quality control (Walsh et al., 2015). Information integration based on RNA-Seq data is most common as a meta-analysis, and it was demonstrated to be valuable in integrating data from different species, tissues and studies [e.g., in (Rau et al., 2014; Sudmant et al., 2015)]. The use of RNA-Seq data can make information integration more manageable and make the data more comparable, as it surpasses the probe effect in microarray data (Rung and Brazma, 2013). However, transcript length bias and nucleotide sequence bias are known challenges that affect the comparability of different RNA-Seq data sets (Rung and Brazma, 2013). Early-stage integration, whenever applicable, is believed to be most suited for comparing two defined growth conditions, yielding a larger number of signature genes. On the contrary, meta-analysis is easier to use in case of largely diverse data sets [reviewed in (Frolova and Obolenska, 2016)]. In addition to the study design and workflow, public gene expression data integration can have its risks. Those include unknown quality, difference in file formats, and difference in experimental setup and conditions (Sielemann et al., 2020).

Re-analysis of the data sets, instead of using summary-level data, is suggested to overcome those risks (Sielemann et al., 2020).

It is possible to consider such data-driven studies as "research parasitism" (Sielemann et al., 2020), however, biological entities can be only seen as part of an ecosystem and with respect to their position in the Tree of Life. Understanding the evolutionary transition in plants and animals and the functional changes in evolved genes can be achieved by integrating information on rarely studied and non-model organisms (e.g., marine protists) from secondary data repositories.

3.1.3 Designing data-intensive applications for biology

The rising availability of high-throughput biological data as well as the potential applications for information integration call for the characteristics of the data system that could support biological data acquisition, storage and analytics. Kleppmann described three concerns when designing data systems: reliability (i.e., working correctly), scalability (i.e., coping with growth in data volume, traffic volume and complexity) and maintainability (i.e., smooth operability, simplicity/abstraction, and evolvability/extensibility) (Kleppmann, 2017). Challenges of reliability and scalability of biological data analytics applications have been discussed (Yang et al., 2017). First, the importance of software reliability is exceptional when analyzing biological data. However, testing and validation are challenged in biological data analytics, in particular due to the gap between the testing data (e.g., simulated data and gold-standard data sets) and the real input data. A possible solution to overcome such a challenge is to, for instance, employ state-of-the-art software testing techniques, such as metamorphic testing, for quality assurance of RNA-Seq expression quantification pipelines (Yang et al., 2017). Second, scalability challenges in handling biological data arise from the large data volumes analyzed and the complexity of analytics. Therefore, approaches to cope with those challenges extend beyond mere parallelism of complex algorithms to include distributed storage and efficient communication in terms of parallel processing and storage. Moreover, as the data load can be unpredictable, on-demand scalable resources with high elasticity, such as cloud computing, could be employed. Finally, utilization of memory-efficient data structures has been also suggested to overcome scalability challenges.

A number of cloud-based RNA-Seq analysis workbenches and application programming interfaces have been developed to promote modularity, scalability and reproducibility. Those include: Oqtans (Sreedharan et al., 2014), MapReduce-based Myrna (Langmead et al., 2010), and lastly Elysium, which supports uniform processing of secondary gene expression data (Lachmann et al., 2018, 2020). Nevertheless, the aforementioned applications have not covered steps of *de novo* assembly of RNA-Seq reads. *De novo* assembly becomes crucial in analyzing organisms whose genomes are yet to be sequenced. A comprehensive guide for the assembly and the analysis of RNA-Seq data on the cloud has been developed (Griffith et al., 2015), providing an

excellent resource on achieving scalability of RNA-Seq data analytics using Trinity, the popular solution for analyzing phytoplankton data (see Chapter 2). Simple workflows could be suitable in case of inapplicability of cloud services (e.g., due to cost or data privacy issues), since they could promote scalability, reproducibility and extensibility.

3.1.4 Case study: Iron stress in photosynthetic organisms

Stress in algae has been reviewed and its definition has been revisited (Fogg, 2001). Stress could be viewed as the change in the environmental conditions that threaten the normal metabolic balance in the organism (i.e., homeostasis), triggering a response to counteract these disturbing effects. An environmental stressor could limit the resources acquisition and/or growth and reproduction in an organism. Algae have the inherited ability to respond to stressors. In addition to grazing pressure and pathogens (Smetacek et al., 2004), algae can suffer from different types of stressors. Those stressors can be categorized into: mechanical (e.g., turbulence), physical (e.g., ultraviolet radiation, osmotic stress, temperature) and nutritional (e.g., nutrient deficiency, pollutants) stressors (Fogg, 2001). It is often the case that algae are subjected to naturally co-occurring stressors. An example of naturally and interacting co-occurring stressors is ice formation in the polar seas that could result in increased salinity, reduced temperature and desiccation [reviewed in (Fogg, 2001)]. Another example is iron limitation, which intersects with other stressors such as low and high light, low copper (as a substitute for iron), and nitrogen [reviewed in (Schoffman et al., 2016)]. In response to stress, algae can move away from the stressor (e.g., high light intensity), alter their metabolism (e.g., limit photosynthesis and cell division), alter their structure (e.g., colony and spores formation), and form symbiotic relationships (e.g., to acquire limiting nutrients) (Fogg, 2001).

A closer look at stress response would suggest that time is an important factor in distinguishing between an inhibiting stressor (on the short-term) and a stimulus (on the long-term) (Fogg, 2001). Borowitzka describes the stages of stress response in microalgae as alarm, regulation, acclimation and adaptation (Borowitzka, 2018). First, when the cell homeostasis is disrupted by the stressor, an alarm response is initiated. Second, cell regulation would occur to restore homeostasis. Third, as cell regulation fails and cellular functions continue to be disrupted, acclimation, which is the change in phenotype (through changes in gene expression), would occur to restore homeostasis. Once acclimation is accomplished and homeostasis is restored, the cells are no longer considered stressed. Forth is adaptation; the change in the genotype of the organism in response to environmental changes. In other words, adaptation can engrave the acclimated phenotype in the cell's genome after the necessary number of generations has been successfully acclimated to the stressful conditions (Borowitzka, 2018). Most of the laboratory investigations study the algal adaptation to non-ecology-driven (i.e., unnatural) single stressors (e.g., batch

cultures design). Recently, adaptation to co-occurring stressors have been successfully demonstrated under laboratory conditions [e.g., ecology-driven iron limitation coupled with ocean acidification (Trimborn et al., 2017; Koch et al., 2019), or change in light and temperature (Strzepek et al., 2019)]. Such experiments are important in understanding the stress response timeline in algae as well as the effects of co-stressors on cellular functions. Transcriptomics offer a rapid cost-effective tool for understanding the timely response of algal species to environmental stressors. Transcriptome functional analysis was used, mainly in diatoms, to identify death markers under chronic stress (Thamatrakoln et al., 2012), monitor stress response at the single-cell level (Shi et al., 2013), and identify and prioritize stressors (i.e., toxic substances) (Osborn and Hook, 2013).

Iron is essential for phytoplankton growth. It serves as an electron carrier in photosynthesis and mitochondrial respiration, and as a cofactor in fatty acid biosynthesis and nitrate metabolism (Marchetti et al., 2012; Harel et al., 2014; Schoffman et al., 2016) (Chapter 2). Therefore, a large portion of the essential gene-set in photosynthetic organisms (Rubin et al., 2015) is iron-dependent (Behnke and LaRoche, 2020). Moreover, iron metabolism genes (e.g., those responsive to iron stress) are not only evolutionary-related (Groussman et al., 2015), but they are also ubiquitous among marine phytoplankton species (Morrissey et al., 2015; Behnke and LaRoche, 2020). Iron stress response has been studied at the molecular level in diatoms (Strzepek and Harrison, 2004; Allen et al., 2008; Lommer et al., 2012), haptophytes (Strzepek et al., 2011, 2012) and cyanobacteria [reviewed in (Morrissey and Bowler, 2012; González et al., 2018)]. The chloroplast has been viewed as a global sensor of environmental stress that results in fluctuations in sugar levels, and triggers metabolic changes (Biswal et al., 2011). Iron is essential to both the photosynthetic and mitochondrial electron transport chains, and markers for iron stress were demonstrated using knockdown experiments in land plants (Vigani et al., 2016). Data integration and co-expression gene network analysis of photosynthetic organisms allowed for deducing conserved gene modules across phytoplankton and land plants (Ferrari et al., 2018). Integrating data and comparing the stress response to iron limitation in a wide range of photosynthetic organisms might reveal core pathways involved in iron metabolism. Moreover, such comparative analysis might shed light on the stages of stress response, and help refining the definition of nutrient limitation. Currently, large databanks of RNA-Seq data on marine algae have become available, the most important of which is Marine Microbial Eukaryote Transcriptome Sequencing Project [MMETSP; (Keeling et al., 2014)]. The MMETSP databank provides data on stress response of a variety of rarely studied algal species. Most relevant, the project has the advantage of employing a unified RNA extraction, sequencing and analysis protocol.

3.1.5 Study objectives

The main **biological objective** of this case study is to comprehensively investigate the gene expression response of a wide range of photosynthetic organisms (namely: diatoms, haptophytes, green algae, cyanobacteria, land plants) to changes in iron availability. Ultimately, the study aims at inferring whether a core response to iron limitation in photosynthetic organisms exists. Moreover, the evolutionary origins of the genes responsible for the core response would be traced.

The **data engineering objective** is to test the usability of a simple pipeline to acquire, integrate and analyze gene expression data archived in public repositories using this case study. Similar designs are discussed below and possible advantages of the presented workflow are highlighted.

3.2 Scalability and information integration

To conduct this data-driven study, a simple **E**xtract-**T**ransform-**L**oad (ETL) workflow for data acquisition, processing, integration, and analysis needs to be designed and developed. Here I designed and implemented a workflow in `Bash` to offer basic functionalities for the case study, which can be easily expanded according to the investigated data sets.

This section addresses: 1) the rationale, challenges and overall requirements of this meta-analysis explaining the data repositories to be curated, 2) the solution implementation steps highlighting the ETL-workflow components, 3) the evaluation and limitations of this case study.

3.2.1 Challenges and project requirements

For the design of this case study, two aspects are to be carefully considered: 1) the type of the transcriptomic data to allow for most manageable and informative integration as well as most comparability of the data, and 2) the analysis plan given the diversity of the species to be included. Therefore, I considered analyzing RNA-Seq data in a meta-analysis fashion (late-stage integration). Choosing raw sequence data as a starting step promotes comparability, as it benefits from employing a unified pre-processing and transcriptome analysis protocol. In addition to the core requirements of a transcriptomic study [described in 2.2, reviewed in (Conesa et al., 2016)], an ETL-workflow for a meta-analysis study would include steps for data curation from secondary data repositories prior to data pre-processing as well as late-stage integration. It is integral to the case study to unify the pre-processing and analysis protocols. Below I list the requirements in an orderly manner.

1. **Data mining and sequence acquisition.** A comprehensive data set of RNA-Seq data from diverse photosynthetic organisms, with focus on iron limitation

and enrichment, is to be curated. The main repositories cover: 1) publications and projects (NCBI PubMed database[1] and iMicrobe[2]), 2) gene expression studies (NCBI Gene Expression Omnibus database; GEO), and 3) sequence repositories (NCBI Sequence Read Archive; SRA and the European Nucleotide Archive; ENA[3]). Domain knowledge is critical in identification and verification of relevant publications and data sets. Afterwards, sequence repositories offer functionalities for seamless data acquisition (e.g., SRA toolkit).

2. **Sequence pre-processing.** The choice of software for reads quality control, filtering low-quality reads and trimming sequencing adapters depends on the sequencing technology used for each data set included in this meta-analysis. The most common sequencing technologies used for marine phytoplankton transcriptomics are: Illumina, SOLiD and Roche 454. Popular quality control software are FastQC (Illumina)[4] and NGSQC (cross-platform) (Dai et al., 2010). For trimming adapters and filtering low-quality reads and bases, Trimmomatic (Illumina) (Bolger et al., 2014) and FASTX-Toolkit (cross-platform)[5] can be used.

3. **Transcript identification and quantification.** There are two considerations for transcript identification for each data set. First, the sequencing technology used dictates the software used. For instance, short reads produced by Illumina (widely used in transcriptomic studies) could be assembled using a large number of assemblers including Trinity, while longer reads from Roche 454 are assembled using the commercial genome assembler Newbler. Second, genome availability can direct the transcript identification strategy towards read mapping against a reference genome rather than *de novo* assembly. *De novo* assembly of paired-end reads from multiple samples (within the same experiment) is encouraged, even in case of available reference genomes (Conesa et al., 2016; Wang and Gribskov, 2017). Transcript (or gene) abundance estimation are currently achieved through alignment-based or alignment-free methods. In an earlier study, Chapter 2, I used an alignment-based method, RSEM, which requires reference transcripts for read alignment. Novel software, such as Salmon (Patro et al., 2017; Srivastava et al., 2020), are used successfully for both alignment-free and alignment-based quantification. Most accurate, and moderately fast, transcript quantification has been obtained through selective alignment against both the target transcriptome and the genome (Srivastava et al., 2020).

[1]National Center for Biotechnology Information (NCBI). Bethesda (MD), National Library of Medicine (US). URL: https://www.ncbi.nlm.nih.gov/

[2]iMicrobe. Youens-Clark K, et al. iMicrobe: Tools and data-driven discovery platform for the microbiome sciences. Gigascience 2019; 8. URL: https://www.imicrobe.us/

[3]The European Nucleotide Archive; ENA. The European Bioinformatics Institute (EMBL-EBI). URL: https://www.ebi.ac.uk/ena/browser/home

[4]Andrews S. FASTQC. A quality control tool for high throughput sequence data. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[5]FASTX-Toolkit. URL: http://hannonlab.cshl.edu/fastx_toolkit

4. **Differential expression, integration and functional profiling.** As in Chapter 2, in a meta-analysis, differential expression analysis is conducted for each data set independently; the results are then combined and interpreted collectively. At that level, information integration is recommended to be through vote count, or through combining ranks, p-values or effect sizes across the included data sets [reviewed in (Frolova and Obolenska, 2016)]. P-value combination is a popular strategy, and it has been used for microalgae transcriptomes meta-analysis (Panahi et al., 2019). As well, its methods (e.g., Fisher method) have been implemented in `R` for RNA-Seq meta-analysis [e.g., metaRNASeq (Rau et al., 2014)].

The need for modularity

In addition to the aforementioned requirements, this case study may lay ground for: 1) construction of draft pan-transcriptomes of the less studied organisms, and 2) investigate the evolutionary origin of the most responsive genes. Therefore, extensibility of the workflow becomes beneficial. A module-based workflow would provide an excellent base for extensibility.

The first prospective aim of this case study is to advance current knowledge on the metabolic potential of non-model organisms through improving *de novo* transcriptome assembly, integrating, at an early stage, transcriptomic data from multiple growth conditions, studies, and sequencing technologies. Hybrid transcriptome assembly from different sequencing technologies is a promising yet challenging strategy for recovering full-length transcripts, as it could yield chimeric transcript contigs. Novel methods for hybrid assembly have been developed [e.g., IDP-denovo (Fu et al., 2018) and rnaSPAdes (Prjibelski et al., 2020)] to combine long and short reads. A modular pipeline can be later extended to offer such hybrid assembly functionality. The assembled contigs can be afterwards checked for chimerism (e.g., by phylogenetic analysis and alignment against published genomes).

The second prospective aim is to extensively characterize the resulting transcripts from the meta-analysis (i.e., meta-genes). In addition to phylogenetic analysis, pathway enrichment analysis and protein-protein interactions prediction would place the common patterns in differentially expressed genes in system-wide context. A modular pipeline could include such analysis in an ad-hoc fashion.

Existing solutions

A pipeline[6] for data curation, pre-processing and analysis has been developed in `Python` for the MMETSP project, which used Illumina sequencing technology. The pipeline uses Trimmomatic (for quality control and adapter trimming), Trinity (for assembly) and Salmon (for transcript quantification). The pipeline was used to generate a transcriptome data set from each biological sample from the project. In the

functional pipeline, modularity is well-established, nevertheless, extensibility requires using `Python` for handling the downstream analysis software. In addition, for the case study, assembly of more than one sample of the target species is sought. In the following subsection, I illustrate a simple workflow in `Bash` that is tailored for addressing the case study requirements.

3.2.2 Solution implementation

In this subsection, I describe the design of the proposed pipeline for this meta-analysis case study and its implementation[7]. Figure 3.1 illustrates the data flow, and the different processes and tools used in the pipeline. The initial implementation presented here addresses transcriptomic data produced by Illumina sequencing technology, the most widely used in marine research.

1. **Data mining.** To compile an input file, the following repositories were manually queried using appropriate corresponding search terms:

   - NCBI GEO DataSets: (iron [All Fields]) AND (Expression profiling by high throughput sequencing [Filter]) AND (txid33090 [Organism:exp] OR txid33634 [Organism:exp] OR txid2763 [Organism:exp] OR txid2830 [Organism:exp] OR Phaeocystis [All Fields]),
   - NCBI PubMed: ((photosynthesis) AND (iron OR Fe) AND (transcriptomic OR rna-seq OR trinity OR RNA/analysis OR Transcriptome/genetics*)),
   - ENA: (phytoplankton OR diatom AND iron), and
   - MMETSP: (iron OR Fe in external_sample_id).

   The search results were aggregated and manually curated into a list of samples from the most relevant experiments (106 samples from 17 experiments on 14 organisms as of March, 2019). Information on sequencing technology, taxonomy and growth conditions were included. The majority of the samples were sequenced using Illumina technology (72 samples).

2. **Data management and sequence acquisition.** The pipeline requires input as a tab-delimited flat file containing parameters for data acquisition (i.e., the unique run identifier for raw data retrieval), (pre-)processing (e.g., sequencing technology and library strand type), and analysis (e.g., experiment identifier, organism name and sample growth condition). The pipeline parses the file and creates directories for each experiment to store raw and processed data as well as the analysis results. The pipeline also creates log files to track progress and report errors. Run IDs are used to download the raw sequence

[6]The Lab for Data Intensive Biology. MMETSP pipeline GitHub repository. URL: https://github.com/dib-lab/dib-MMETSP

[7]The pipeline is available upon request through GitHub.

data in FASTQ format using `parallel fastq-dump` from NCBI SRA using SRA toolkit (v2.9.2).

3. **Sequence pre-processing.** As in 2.2, the pipeline uses Trimmomatic (v0.32) to trim sequencing adapters and to eliminate low-quality bases and very short reads in each run file.

4. **Transcript identification and quantification.** Quality-filtered reads from each experiment are pooled into a single input file for one assembly run by Trinity (v2.0.4). For transcript quantification, I use Salmon (v1.0.0) in selective alignment-based mode; the newly developed algorithm offering more accurate transcript abundance estimation (Srivastava et al., 2020).

5. **Differential expression, integration and functional profiling.** Differential expression is conducted at transcript-level through pairwise comparisons of the sample types in an experiment using DESeq2. As in 2.2, only transcript of length ≥ 300 bases and sum of rounded counts ≥ 40 are included. The following default cutoffs are considered: 1) false discovery rate (FDR) ≤ 0.001 and 2) absolute logarithmic fold-change (LFC) ≥ 2. The differentially expressed contigs are characterized as in 2.2 using TransDecoder (v2.0.1) and Trinotate (v2.0), for coding sequence prediction and annotation, respectively. The differential expression results are integrated for the contigs of known function, and a combined p-value is reported for each differentially expressed gene.

3.2.3 Evaluation and critical appraisal

Meta-analysis of transcriptomic data represents a case of the reuse of public research data that can help inferring patterns in gene expression across different tissues and organisms. A scalable solution for data curation and processing is necessary to successfully conduct such data-driven studies. I designed and implemented a pipeline to promote scalability and modularity for RNA-Seq data meta-analysis. This pipeline could be used to comprehensively investigate the gene expression response of a wide range of photosynthetic organisms ($n = 106$). The usability of the initial implementation of the workflow could be evaluated with respect to data acquisition, data pre-processing and analysis, and information (late-stage) integration by analyzing a selected subset of well-studied species ($n = 5$) sequenced using Illumina sequencing technology (*Synechocystis* sp. PCC 6803, *Chlamydomonas reinhardtii*, *Chaetoceros debilis*, *Thalassiosira oceanica*, and *Oryza sativa*).

The ultimate goal of the FAIR Guiding Principles for scientific data management and stewardship (Wilkinson et al., 2016) is to optimize the reuse of data, promoting data as findable, accessible and interoperable. This case study sheds light on technical and research-relevant concerns when analyzing public environmental data.

Regarding the research-relevant concerns, the evaluation of the workflow using real data was not concluded. In spite of data findability, accessibility and integrability,

a major growing concern became data ownership; whether a meta-analysis might conflict with the prospective plans of the data owners, consortia or individuals. Therefore, personal communication with data owner to seek approval would be courteous. In addition, publishing the meta-analysis-based study results in peer-reviewed journals would ensure the novelty of the question and the quality and comprehensiveness of the data sets, challenging the perception towards the reuse of data.

On the technical side, the case study and the accompanying workflow have simple requirements. Nevertheless, more sophisticated data-driven cases and/or analyses are of interest, and the modularity of the workflow would become essential in such cases. First, concerning transcript identification and quantification, a particular study might require hybrid transcriptome assembly of reads from different sequencing technologies. For that, chimeric sequences can be minimized by filtering data based on studies' quality, and by employing a number of assembly quality assessment tools [e.g., DETONATE (Li et al., 2014a)] and protocols[8]. Second, concerning information integration, modularity would facilitate experimentation with sophisticated analyses. For instance, it has been considered to evaluate the reduction in transcriptome diversity utilizing Shannon entropy as a function of increased stress levels, providing a way of prioritizing stressors in phytoplankton. Analysis of both simulated and real data using a modular extensible workflow would allow experimenting with, among others, information theory-based analysis. Third, a further step towards scalability would be cloud deployment, for example on Amazon Web Services (which offers a genomics analysis solution[9]) or on iMicrobe. Such elasticity would help accommodate the unpredictable data load.

3.3 Concluding remarks

Cell transcriptome is a very sensitive proxy for the amount of change in environmental conditions affecting the cell. Gene expression profiling technologies yield large amounts of transcriptomic data available in public repositories, challenged by data acquisition (i.e., transfer, storage and harmonization) and analytics. Both the challenges and the value of the integrated information motivate the development of effective data handling approaches to decipher and evaluate the combined biological knowledge.

This chapter presents a case of meta-analysis of transcriptomic data to infer a potential core response to iron limitation in photosynthetic organisms. The case study highlights a few bottlenecks in secondary sequence data handling, namely: 1) a scalable approach for data acquisition and transformation, 2) a modular workflow for

[8]Trinity RNA-Seq Wiki. URL: https://github.com/trinityrnaseq/trinityrnaseq/wiki/Transcriptome-Assembly-Quality-Assessmen

[9]Amazon Web Services: Genomics. URL: https://aws.amazon.com/health/genomics/

experimentation with hybrid assemblies and sophisticated analyses, and 3) a flexible computing plan (e.g., cloud computing). The simple pipeline presented here would promote scalability and modularity for RNA-Seq data meta-analysis. As it is the case with the reuse of public data, evaluating this workflow using public data would be possible through communication with the data owners to address conflict of interest.

Even though sequence data are the most abundant and computationally intensive biological data, structured relational data could also present an opportunity and a challenge for data processing and analysis. Scalability equally challenges transforming and analyzing secondary non-sequence data for large-scale studies for outcome prediction, which I address in Chapter 4.

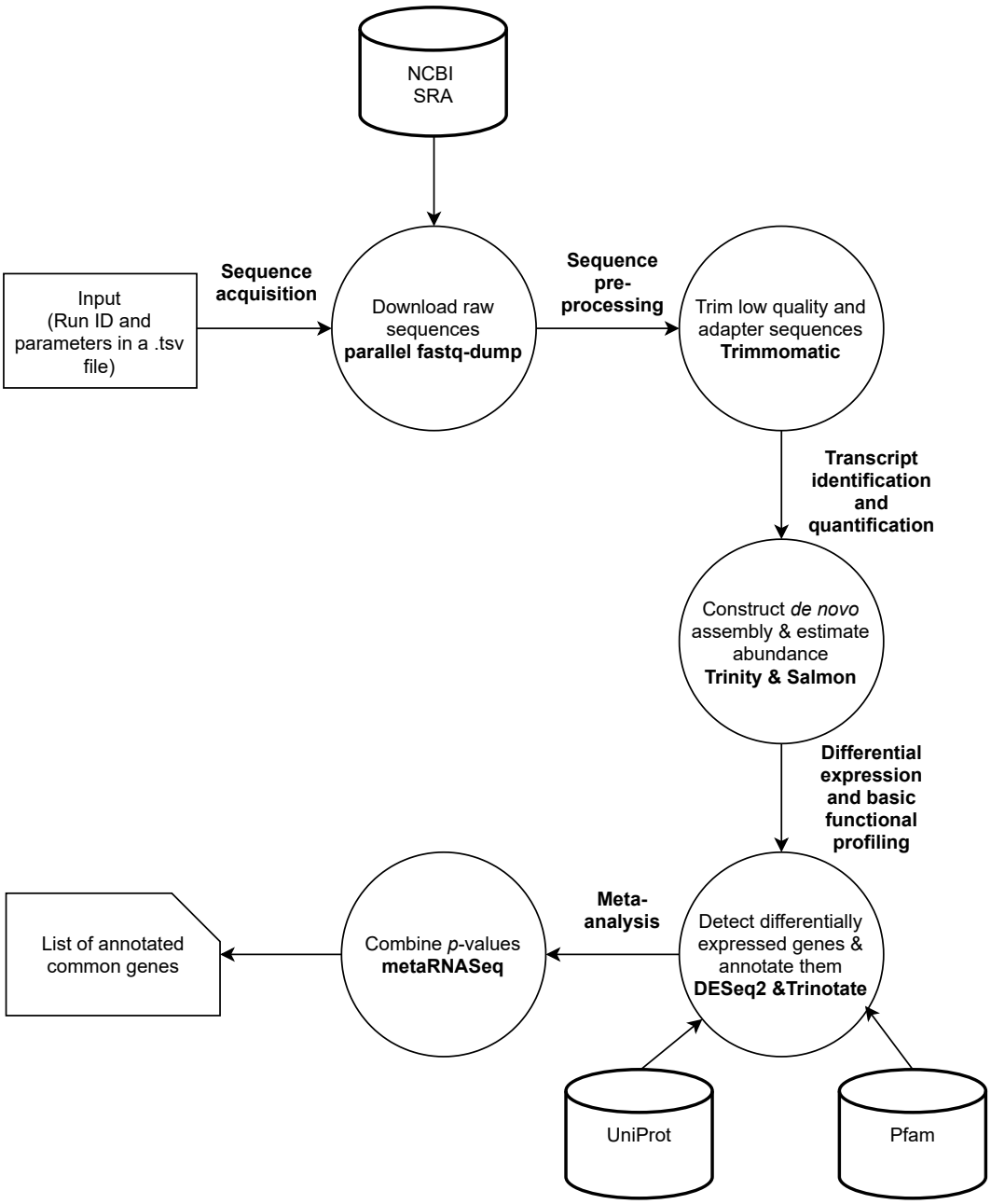


FIGURE 3.1: A data and processes flow diagram for RNA-Seq data meta-analysis. It illustrates the data acquisition, pre-processing and analysis processes for the case study. In each box, the process (upper) and the tool (lower) are stated. The rectangular boxes represent input data (a tab-delimited flat file); the circles represent the processes and the card shape represents the output. The arrows illustrate the direction of data movement. The repositories queried for sequence data and used for sequence functional analysis are represented as well. NCBI SRA = NCBI Sequence Read Archive, UniProt = The universal protein knowledgebase, Pfam = The Pfam protein families database.

Chapter 4

Integration and Statistical Modeling of High-Dimensional Data: A Case of Secondary Health Data

Adverse drug events (ADEs) represent a burden on the health care system as they cause significant morbidity and mortality. In Europe alone, 3-10 % of all hospital admissions are due to ADEs (European Commission, 2008). Drug safety studies attempt at evaluating drug effectiveness and minimizing the occurrence or severity of an ADE, specially rare adverse events. A patient's response to a drug is the sum of many factors including genetics, nutrition, alcohol consumption and smoking, co-morbidities, and concomitant drug use. Being able to successfully predict risk in patients and to identify the characteristics (e.g., drugs and diseases) that lead to an increased risk to suffer an ADE is of utmost importance. Two data sources are typically used for detecting ADEs: spontaneous reporting systems and longitudinal pharmacoepidemiological databases. Spontaneous reporting systems are limited to the drug in question and the ADE itself. To account for the many factors leading to an ADE, routinely collected longitudinal health care data could be a valuable source.

In the past few decades, a steep rise in routinely collected health data sources, referred to as electronic health care databases, has taken place. These repositories represent a readily available, cheap and fast source of data that can be used to monitor drug safety in large populations in the post-marketing phase. In the genomics era, drug safety studies would benefit from analyzing such data in the light of molecular biology. This requires: 1) extraction and integration of molecular biology-related ontologies from public knowledge bases, and 2) utilization and development of genomics-relevant statistical methods, which, in turn, require data transformation and scalability for studies on large populations.

This chapter addresses the aspects of data integration, and scalability of current

implementations of specialized statistical methods using a case study from pharmacoepidemiological research as an example of secondary routinely collected data. This case study attempts at predicting the risk of ADEs mostly seen in patients using anticoagulants given the patients' drug and disease profiles in cases and matched controls using data from the German Pharmacoepidemiological Research (GePaRD). The study compares predictions from various methods that do and do not incorporate knowledge on drug and disease molecular pathways. The basic methodology proposed in this case study, to which I contributed, has been sketched and published (Appendix B.2; in German). A few methods were screened and tested on simulated data (Appendix A.2). The work featured in this chapter guided the improvement of the study design and methods selection, and resulted in a draft manuscript to be considered for submission to *Drug Safety* journal (Appendix B.3). The project is funded by the Innovation Fund of the German Joint Federal Committee (G-BA, 01VSF16020). I thank the statutory health insurance provider, which provided the data used for this case study, Die Techniker (TK).

Using the aforementioned case study, I investigate and use the best solutions to extract-transform-load data from public knowledge bases and from GePaRD, and apply a number of known and novel statistical methods in `R` and `Python`. In addition, I compare the predictions of these models. The chapter also highlights scalability issues in data acquisition and analysis, data security, and the challenges of data-driven simulation studies.

The chapter is structured as follows: Section 4.1 gives background information on electronic health care databases and their usefulness in ADEs detection, the statistical methods used, the data source, GePaRD, and an overview on the specific case study. Section 4.2 addresses the chapter objective in terms of project requirements, solution implementation and results evaluation. Concluding remarks are presented in the last section.

4.1 Background

4.1.1 Electronic health care databases

The past few decades witnessed a steep rise in routinely collected health data sources. These sources, referred to as electronic health care databases (EHDs), span three categories, namely: record linkage systems starting in the 1960s (e.g., national disease and death registries), electronic medical records and health care claims databases (Pacurariu et al., 2018). Electronic medical records represent the most common form of EHDs in Europe (Pacurariu et al., 2018); they provide detailed information on patient's symptoms, and medical examinations and their results. Health care claims databases store routinely collected data for reimbursement purposes by statutory health insurances (SHIs). They contain demographic information (e.g., age, sex,

occupation), prescription information (e.g., drug name, dose, duration, and possibly route of administration and therapeutic indication), and diagnosis information (e.g., in- and outpatient diagnoses and procedures) (Schneeweiss and Avorn, 2005; Pacurariu et al., 2018). An example of health care claims databases is the German Pharmacoepidemiological Research Database (GePaRD) (Pigeot and Ahrens, 2008).

The strengths of the EHDs, such as large size, realistic representation of the population, and availability at a low cost, qualify them for several applications (Schneeweiss and Avorn, 2005; Pigeot and Ahrens, 2008). These applications can be classified into drug-related, health policy-related, and data usability and validation-related applications. First, drug-related applications include drug utilization studies (i.e., the number of prevalent and incident users of a particular drug or biosimilars), and drug safety and effectiveness studies in the post-marketing phase, specially in populations that are not included in clinical trials (e.g., elderly, children, and pregnant women). Second, health policy applications include studies of patterns in physician prescription practices, and studies of different drug reimbursement policies and their effect on health outcomes (Schneeweiss and Avorn, 2005). Third, data usability and validation studies include quality control, validation of diagnosis coding and development of drug utilization algorithms. In the era of machine learning and genomics, applications of EHDs are currently encompassing, for example, respectively, text- (McTaggart et al., 2018) and data mining (Umemoto et al., 2019), and merging with genomic data (Hall et al., 2016).

There is a number of considerations regarding EHDs management and analytics. EHDs are often used in conjugation with other data sources. For instance, data are required to be validated perhaps through linkage to national disease and death registries, while diagnosis and prescription data could be transformed using ontologies such as the international classifications of diseases and drugs. Due to the nature of EHDs data, high standards are to be employed for analyzing EHDs in epidemiological studies. For instance, proper study designs (i.e., decision on a cohort, case-control or nested case-control design, definition of new users as exposed individuals, and identification of and controlling for confounding variables) and rigorous statistical tests are required. The complexity of both data management and analysis increases in case of integrating data from multiple EHDs (Schneeweiss and Avorn, 2005; Andrews et al., 2014).

To overcome such complexity, using multiple data sources and/or in conjugation with genomic and molecular data would require: 1) development of a unified data model, 2) utilization of ETL workflow or alternatives to populate the model with the data, and 3) deployment of analytical workflows on high-performance computing resources for such large-scale studies (Curcin et al., 2008). From an analytics perspective, standard pharmacoepidemiological methods fall short in analyzing big data, therefore, the alternatives include applying data mining and machine learning methods in analyzing EHDs, and adapting standard epidemiological methods for

high-performance computing platforms. Moreover, from an information extraction perspective, analyzing data from EHDs incorporating prior knowledge on drugs similarity and drug targets requires development of novel scalable epidemiological methods.

4.1.2 Data-driven methods in pharmacovigilance

Pharmacovigilance (PV) is one of the most central aims for using EHDs. The World Health Organization (WHO) defines PV as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem" (World Health Organization, 2021b), including data gathering activities (Lu, 2009). Following its release, a drug's safety during the post-marketing phase can be assessed through post-authorization safety studies (PASS) or post-marketing surveillance studies. These allow to evaluate drug effectiveness and minimize the occurrence or severity of an ADE, particularly rare adverse events (Pacurariu et al., 2018). A "signal" of an ADE (i.e., a possible relationship between an ADE and a drug) can be detected in two typical data sources: spontaneous reporting systems and longitudinal pharmacoepidemiological databases (Suling and Pigeot, 2012). An example of a spontaneous reporting system is FAERS[1].

As a data source, spontaneous reporting systems can only be used to answer questions related to patients that reported taking a particular drug and suffering ADEs. As mentioned earlier, longitudinal pharmacoepidemiological databases provide additional information such as co-administered drugs, co-morbidities and demographic variables (Suling and Pigeot, 2012) on all individuals that were prescribed a particular drug, and those who experienced the ADE.

While choosing the data source is important, choosing the signal detection method is as important. A signal detection method of choice should be able to utilize available variables (i.e., demographic variables, co-medications and co-morbidities), and scalable to handle large number of subjects and variables of several types. Throughout the years, a wide range of data-driven methods were developed and used in PV, some of which were applied to EHDs. These data-driven methods include: 1) data mining-based prediction of ADEs; reviewed in (Harpaz et al., 2012; Suling et al., 2013), 2) prediction using a combination of data mining and/or machine learning methods, and molecular similarity between drugs (Vilar et al., 2012), molecular pathways (i.e., drug targets) (Liu et al., 2012), drug-drug interactions (Liu et al., 2017), and clinical coding (McMaster et al., 2019), 3) detection of rare ADEs using standard (Chan et al., 2015) and machine learning-based methods, 4) ADE detection using different ontologies (Saunders et al., 2005; Winnenburg et al., 2015), and 5) causal inference (Schneeweiss, 2018). Basic study design recommendations for ADE signal detection in EHDs could be found in (Schneeweiss, 2010).

[1]FDA Adverse Event Reporting System (FAERS). URL: https://www.fda.gov/drugs/drug-approvals-and-databases/fda-adverse-event-reporting-system-faers

4.1.3 The German Pharmacoepidemiological Research Database

GePaRD, the data source for this case study, is a health care claims database in Germany, established in 2004 by the Leibniz Institute for Prevention Research and Epidemiology - BIPS (Pigeot and Ahrens, 2008). GePaRD currently contains claims data obtained from SHIs on more than 24 million insured individuals of the years 2004 to 2016. GePaRD covers patient information divided into four data dimensions: socio-demographic data, inpatient diagnoses data, outpatient (ambulatory) diagnoses data, therapeutic and diagnostic procedures on a quarterly basis, and outpatient drug dispensation data. These data dimensions are represented by the colors in Figure 4.1. Diagnoses are coded according to the International Classification of Diseases, 10th revision, German Modification (ICD-10-GM). Outpatient therapeutic and diagnostic procedures are coded according to the German procedure classification system for surgical and medical procedures (OPS). Drug dispensations can be linked to a reference database via the central pharmaceutical reference number (PZN), and drugs can be mapped to the Anatomical Therapeutic Chemical Classification System (ATC). For each drug, the reference database contains up-to-date information on active substances, brand names, strengths, dosage forms, and defined daily doses. The data dimensions are linked by a pseudonymous subject identifier (ID). GePaRD is stored in a relational database management system (ORACLE®) maintained by BIPS. The unified data model for GePaRD was developed by BIPS for integration of data from different SHIs, imposing strict data transfer, protection and quality measures. GePaRD is regularly updated (extended with patient time) providing concurrent information on a large fraction of the German population (Pigeot and Ahrens, 2008).

In Germany, social security data (including administrative health care data) are protected by Article 75, Social Code Book (SGB) X (Pigeot and Ahrens, 2008). This article permits, under rigid constraints, the use of data for scientific research purposes without the need for the informed consent of each insurant. Therefore, each study based on GePaRD, requires both the participating SHIs approvals and the approvals of the corresponding regulatory regional or nationwide authorities for data use (Enders, 2017). Such high-quality large-scale structured database is perfectly suited for: 1) use in large-scale drug safety and drug utilization studies on free-living populations (Pigeot and Ahrens, 2008), 2) PV and concurrent drug monitoring, 3) identifying trends in prescription, and 4) conducting siblings and familial linkage studies (as insurance data can link mothers and children).

4.1.4 Case study: Predicting patient risk for adverse drug events in health care claims data using functional targets knowledge

A patient's response to a drug, including susceptibility to ADEs, is the sum of many factors such as: genetic makeup (Meyer, 2000; Phillips et al., 2001), microbiome

(Rizkallah et al., 2010), lifestyle, e.g., nutrition, alcohol consumption and smoking (Alomar, 2014), co-morbidities (Dumbreck et al., 2015), and concomitant drug use (Stewart et al., 2017). Specially in elderly patients, polypharmacy and multimorbidity can lead to an increased risk of ADEs (Dumbreck et al., 2015; Schöttker et al., 2017). It is, therefore, important to successfully predict the risk and identify the characteristics of patients that lead to suffering an ADE.

Drugs and diseases combinations can increase the susceptibility of a patient to a particular ADE through their interactions. A basic case is that concomitant use of similar active drugs can augment the drugs intended effect (e.g., hypoglycemia as a result of multiple active antidiabetic agents). In a less apparent case, ADEs occur due to the drugs effect on off-targets (i.e., unintended targets) (Lounkine et al., 2012). For example, the withdrawn synthetic estrogen chlorotrianisene inhibits COX-1 enzyme and inhibits platelet aggregation, thus it can exacerbate bleeding in a patient taking anticoagulants (Lounkine et al., 2012). Another example is the use of antidepressants and antipsychotics simultaneously. Both drug groups block muscarinic receptors, and if combined, their synergistic effect on off-targets can lead to urinary retention as an ADE. Underlying co-morbidities can affect drug choice and ADEs. For example, selective serotonin reuptake inhibitors would increase the risk of bleeding in depressed patients with myocardial infarction and thus drives drug choice modification (Dumbreck et al., 2015). Therefore, to mitigate the risk of an ADE of a particular drug and thus to identify possible actions, such risk factors and the interaction between them need to be identified.

In the past, spontaneous reporting systems were preferably used for signal detection. However, longitudinal databases currently offer a more comprehensive source of information that are necessary to better assess individuals risk of an ADE. In particular, longitudinal databases provide additional information on co-administered drugs and operations, co-morbidities in addition to important confounding variables (e.g., age and sex) (Suling and Pigeot, 2012). Current signal detection methods were originally developed for spontaneous reporting systems, and they continued to be used for analyzing longitudinal data (Arnaud et al., 2017). A limitation of this approach is the need for transforming longitudinal data into a spontaneous report (create pseudo-reports of drug exposure and events) (Arnaud et al., 2017). Consequently, this approach disregards co-medications and co-morbidities (Suling and Pigeot, 2012). An alternative approach for signal detection in longitudinal data is to utilize traditional pharmacoepidemiological study designs (e.g., matched case-control and self-controlled designs) combined with statistical methods that can adjust for confounders and handle a large number of predictors (Arnaud et al., 2017).

In Figure 4.2a the classical approach to predict patient risk in PV is schematically represented. Classically, patient risk is predicted based on the associations between individual risk factors (drugs and diseases) and the ADE. This approach is limited by 1) the restricted available information for patients exposed to these drugs, and

2) the scale at which statistical models can handle and utilize such relatively large number of variables. Methods for variable selection (e.g., penalized logistic regression) are important data dimensionality reduction approaches for large-scale signal detection studies. Another approach to reduce data dimensionality is testing for association between a group of covariates (e.g., drugs/diseases) and a specific outcome (e.g., ADE). This concept is well-established in genetic epidemiology (shown schematically in Figure 4.2b). In genetic epidemiology, pathway analysis approaches allow for combining evidence for associations between single covariates (e.g., genes) and the outcome (e.g., phenotype), which 1) leads to better signal detection (Yu et al., 2009), and 2) helps interpreting the risk factors according to their biological pathways. Pathway selection methods include Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005; Mooney et al., 2014). GSEA is suitable for testing associations between the phenotype and a group of single nucleotide polymorphisms or genes in particular pathways. However, GSEA is not the method of choice for outcome prediction. Recently, group-based penalized regression (Friedman et al., 2010) became widely applied in genetic epidemiology to infer associations while incorporating pathway (i.e., group) information and predict biological outcomes (Breheny and Huang, 2009; Breheny, 2015). Overlapping group logistic regression facilitates handling overlapping pathways in regression models (Zeng and Breheny, 2016).

Similarly, we propose that instead of assessing the associations between the drugs and diseases, and the ADE directly, as in Figure 4.2a, the associations between the groups and the ADE shown should be investigated. We propose that drugs and diseases are grouped by the functional targets (FTs) they interact with. An FT is a pathway of interacting biomolecules (e.g., enzymes, receptors) that are affected by the drug (Overington et al., 2006) or associated with a disease. Online databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2017) and Therapeutic Target Database (TTD) (Li et al., 2018) can be queried to curate the FTs. By exploiting domain knowledge to assign drugs/diseases to groups, we can possibly improve risk prediction by increasing the power for detecting associations. In addition, pooling the data within each group reduces data dimensionality. Moreover, target-based prediction of ADEs might help resolve target pathways (more importantly unintended target pathways) of drugs, which can better explain the underlying mechanisms of ADEs.

4.1.5 Study objectives

Methodological/biological objectives: The methodological and biological aims of the study are: 1) to compare classical PV methods and GSEA-based methods in their ability to predict the risk of ADEs in longitudinal data, and 2) to investigate the effect of grouping covariates based on their functional targets on individual outcome and, whenever applicable, group predictions.

Data engineering objective: The data engineering aim is to optimize and develop scalable portable solutions for: 1) transformation of structured high-dimensional data as health care claims data in conjugation with molecular knowledge from online data sources, and 2) adapting statistical methods for high-dimensional data and high-performance computing resources. A number of optimizations are considered including parallelization as well as utilization of relational databases and on-disk intermediate storage.

4.2 Integration and statistical modeling of high-dimensional data

This chapter presents an example of the need for solutions for data transformation and integration required for knowledge transfer (from genomic knowledge bases) and method transfer (from genetic epidemiology), and for adapting statistical methods to large-scale studies on longitudinal data. In particular, this section addresses: 1) the challenge and requirements of this case study, the data to be integrated and transformed, and the statistical models to be adapted for scalability, 2) the solution implementation steps, 3) the evaluation of data preparation and analysis, and 4) the limitations of this case study and of the solution implementation.

4.2.1 Challenges and project requirements

To achieve the methodological and therefore the biological aims of the study, four components are required: 1) a grouping structure for the covariates (i.e., drugs and diseases), 2) patient data extracted from longitudinal databases following an appropriate epidemiological study design, 3) statistical methods that support prediction, and 4) the outcome to be predicted (i.e., ADE). Handling the grouping structure (i.e., FT) data and the patient data (i.e., longitudinal GePaRD data) presents a classical ETL case, as each of these two components requires extraction, transformation and integration. Below, I describe each of the components in detail.

1. **The outcome (i.e., the ADE):** In early 2010s, a number of novel anticoagulant drugs were released, named non-vitamin K (or novel) oral anticoagulants (NOACs). NOACs were successfully used for treating or preventing blood clots. Patients benefiting from these drugs might have atrial fibrillation and NOACs lower the risk of stroke caused by blood clots. They might be undergoing a hip/knee replacement, where NOACs can lower the risk of formed blood clots in the legs (deep vein thrombosis) or in the lungs (pulmonary embolism). Patients also might be at risk of stroke, heart attack, or other cardiovascular problems. However, these novel anticoagulants may lead to bleeding events in the gut (gastrointestinal bleeding and ulcers) and in the brain (intracranial bleeding). NOACs might also induce liver toxicity. In this case study, we use various statistical methods to detect the signals of the aforementioned ADEs in

the German population using longitudinal data. The methods would predict the ADE risk, and, whenever applicable, select the risk factors (i.e., age, sex, NOACs and/or other drugs and diseases) for suffering these ADEs. In the chapter, I discuss only the first ADE of interest namely gastrointestinal bleeding and ulcers.

2. **Study design based on GePaRD:** A suitable epidemiological study design is integral to signal detection in longitudinal data. Two study designs were considered for this case study: cohort study and matched case-control study. In a cohort study design, we would use demographic, diagnosis and dispensation data available on the whole German population in GePaRD in a specified time period starting from when NOACs were released (e.g., 2010-2016). In a matched case-control study design, we would use the data available for cases (patients that were diagnosed with the ADEs), and match those cases with controls of the same birth year and sex. Advantages of using a matched case-control design are: 1) to reduce data dimensionality (i.e., both the number of observations and covariates) and 2) to balance the patient time (and therefore the covariates) between cases and controls (i.e., cohort exit date for the matched controls and the cases becomes the index date on which the ADE was diagnosed in the case). For a comprehensive investigation, as risk factors, we considered both in- and outpatient diagnoses, in addition to drugs, which is expected to increase data dimensionality. Test runs showed that a large 6-year cohort study is not feasible within the current infrastructure, discussed below. Therefore, I used a matched case-control study design.

3. **Functional target data:** Figure 4.3 gives an overview on the course of the proposed analysis categorizing SHIs data according to biologically relevant FTs and testing for associations with the ADE of interest. The figure also illustrates how drugs/diseases are to be related to FTs. As drugs and diseases (i.e., predictors) are required to be grouped in a biologically relevant manner, target genes and target pathways were considered. Most drugs and diseases have one or more known biological FTs (e.g., a receptor, an enzyme, an ion channel that is expressed by a gene), and those targets are naturally grouped in pathways. Pathways provide a high-level classification of drugs and diseases and, therefore, reduce data dimensionality. To acquire such information, three knowledge bases were considered: STITCH[2] (Szklarczyk et al., 2016), KEGG[3] and TTD[4]; each provides a facet of the FTs (described in Figure 4.3). The facets that would be used by the statistical tests are drug-disease (indication), drug-target and disease-target interactions. Those facets are possible to curate from KEGG and TTD, where TTD, a manually curated database, has several advantages over KEGG. First, TTD provides clear comprehensive information on drug-disease relationships according to the ICD coding system. KEGG, in contrast, is highly selective as it provides information on the diseases

that primarily have an underlying genetic cause. Second, TTD can be readily cross-referenced with the well-known knowledge base KEGG, and with ICD and ATC systems (used in GePaRD and SHIs data). Therefore, I considered using TTD for curating grouping structures.

4. **Statistical modeling and prediction:** As aforementioned, a statistical method of choice should be scalable for handling and utilizing a large diverse number of covariates, and suited for prediction. Moreover, to achieve the methodological aims, the method should be able to handle data as blocks/groups of FTs. All appropriate methods should be compared with respect to their ability to 1) infer which groups (i.e., FTs) leading to an increasing risk of experiencing a certain ADE, and 2) predict whether a patient might experience the ADE given his/her drug exposures and comorbidities. Methods that do not include prior knowledge of blocks structure are also required to investigate the effect of grouping on the prediction.

   We screened and tested a number of methods on simulated data, see the results in Appendix A.2. The methods were then filtered based on subject knowledge and scalability. The methods considered for this case study are based on penalized regression [the lasso (Tibshirani, 1996) and the overlapping group lasso (Friedman et al., 2010; Zeng and Breheny, 2016)], machine learning [block forests (Hornung and Wright, 2019)] and a GSEA [the adaptive rank truncated product; ARTP (Yu et al., 2009)]. As in Yu *et al.*, the ARTP method has not been developed or used for prediction (Yu et al., 2009). We, therefore, developed an implementation of the ARTP that can be used for prediction. This implementation was introduced and tested on simulated data (Appendix A.2), where it showed a superior performance compared to other tested methods with respect to both prediction and inference of associated groups, specially in case of a weak signal. Group lasso, block forests and the new implementation of the ARTP incorporate group information in predicting the outcome in contrast to the lasso. Therefore, these group/block-based methods are compared to 1) standard logistic regression model (Cox, 1958), 2) a penalized regression method that does not include prior knowledge of group structure (e.g., the lasso), and 3) a penalized regression method that does, which we refer to as the naïve group lasso (NGL). The NGL is based on creating a group variable that is either the sum of the covariates within that group or a value of 1 when any of the group covariates is 1, and 0 otherwise. The results of the simulation study showed that sum-based NGL had better predictability compared to occurrence-based NGL (Appendix A.2), therefore group size weighted sum-based NGL is considered here. Finally, the performance of the methods is compared in terms of recall, precision and F1-score.

[2]STITCH: Search Tool for Interactions of Chemicals. URL: http://stitch.embl.de/
[3]KEGG: Kyoto Encyclopedia of Genes and Genomes. URL: https://www.kegg.jp/
[4]TTD: Therapeutic Target Database. URL: http://db.idrblab.net/ttd/

Existing solutions and their constraints: Time and memory

The project funding requires the solution to be available for distribution among the project partners (e.g., the SHIs involved). Therefore, cross-platform solutions were considered. `R` is a strong platform for both data acquisition and analytics. `R` can also effectively handle the diverse ecosystem of data sources considered for this study (e.g., TTD flat files and GePaRD data stored in a centralized relational database management system). For portability and distribution, an `R` package could be developed to contain all data management and analysis processes.

Despite `R`'s strengths, there are known performance and scalability limitations [see (Morandat et al., 2012; Wickham, 2014)]. In particular, inefficient memory utilization and lower computation efficiency are highly relevant to data engineering when handling high-dimensional data. As data grow in size and complexity, these aspects eventually lead to two classes of runtime errors: out-of-memory (OOM) and prohibitive execution time (exceeded walltime limit). At runtime, `R` operations on high-dimensional data create intermediate data objects, which usually grow in size and dimensions. This expansion in object size hinders efficient computation on modern multi-core CPUs, as the data used in calculations can be barely contained in the CPU caches or even main memory, a problem augmented by memory fragmentation resulting from constant objects growth and relocation (Burns, 2011). On the one hand, the `R` program can be abruptly killed by the kernel's OOM killer service due to main memory capacity exhaustion. On the other hand, the reduced efficiency might cause the operations to be extremely slow and exceed execution time limits.

Throughout the years, alternatives have been proposed to improve these limitations in `R`, such as new implementation of the language [reviewed in (Wickham, 2014)]. Improvements can be also achieved through code optimization (e.g., parallelization) and utilization of object classes that are high-dimensional data-friendly (e.g., `data.table` and sparse matrix).

Application field constraints: Data dimensions and data protection

In pharmacoepidemiological studies, data sets from longitudinal data are often of lower dimensions than that intended for this case study. For instance, in (Pisa et al., 2019), a large matched case-control study of 16,750 cases and 1,673,320 controls, and approximately 50 covariates was analyzed using a logistic regression model. Current implementations of statistical models (e.g., logistic and penalized logistic regression) are rarely used to analyze larger data sets. To be able to use those methods and others such as block forests and the ARTP, data dimensionality is a constraint. Setbacks are expected when it comes to analytics of such high-dimensional data in terms of both software implementations and hardware constraints. In particular, to comply with data protection constraints, data acquisition, transformation and analysis operations of GePaRD data, including this case study, are required to be run on

the institute's computing cluster, which, even though powerful, is a limited shared resource.

4.2.2 Solution implementation

Throughout the development phase, OOM and exceeded execution time problems arose. In particular, in case of a cohort study design, data acquisition and transformation was repeatedly halted by OOM issues, such that analytics testing was not possible. In the matched case-control study design, both transformation and analytics experienced OOM and exceeded execution time issues. It meant that even with the dimensionality reduction by adopting a matched case-control study design, resources and time represented constraints.

Resources availability and time are important aspects of signal detection studies; therefore, the study's aims were translated into finding a statistical method that produces the most accurate predictions while using reasonable resources. In this subsection, I describe the study population and data dimensions of the matched case-control study, data acquisition, transformation and analysis for each of the project components. I focus on the successfully applied implementations that did not experience runtime errors.

Machine specifications The solution runs on a 28-node cluster operated by CentOS Linux 7, with an Intel® Xeon® CPU E5649 @ 2.53GHz 2 x 6-core CPUs, 12MB shared cache, 98GB of RAM per node, and 10TB disk space; R version 3.4.3 (2017-11-30).

1. **The outcome:** Gastrointestinal bleeding was chosen as the ADE of interest. Incidence of gastrointestinal bleeding was defined based on diagnoses of gastric ulcer (K25.0, .2, .4, .6), duodenal ulcer (K26.0, .2, .4, .6), peptic ulcer (K27.0, .2, .4, .6), gastrojejunal ulcer (K28.0, .2, .4, .6), or gastritis and duodenitis (K29.0), according to ICD-10-GM.
2. **Longitudinal data (GePaRD):**
    - *Study population and design:* In this case study, a matched case-control study design was applied. The study entry date was between 1 January 2015 and 31 December 2016. Data from the Techniker Krankenkasse (TK) were used. Eligible subjects had to be 1) aged 18 years or older at the time of cohort entry (birth year $\leq$ 1996), 2) continuously insured throughout the study period (with up to 3 days gap), and 3) living in Germany. Eligible subjects that had a valid date of death (i.e., on the basis of hospital discharge cause and date, and end of insurance case and date) were retained in the cohort as long as the ADE occurred before the date of death. Subjects for whom year of birth and sex were not available were not considered for study entry.

- *Case definition and matching protocol:* Cases (i.e., subjects suffering from the ADE of interest) were identified on the basis of in- and/or outpatient diagnoses, whichever available. Inpatient diagnosis must be either main or other main discharge diagnosis. Outpatient diagnosis must be an assured diagnosis. Cases must not have the ADE of interest during the baseline period (July 1, 2014 - December 31, 2014) and the first quarter of 2015 (January 1 - April 30, 2015). Controls must not have the ADE of interest throughout the baseline and the study periods. Cases are not eligible to be controls at any time. Matching [1:10 without replacement (Robins et al., 1986; Pearce, 2016)] based on sex, year of birth and index date was performed. Cases and eligible controls were divided by identification number into a training data set (even numbers; for fitting) and a test data set (odd numbers; for prediction) before the matching was performed.
- *Predictors definition:* Both co-morbidities and concomitant drug administration were considered as risk factors (i.e., predictors) for the ADE of interest. Co-morbidities were obtained from in- and outpatient diagnoses (main and other main discharge diagnosis, secondary and auxiliary diagnosis and diagnosis for ambulatory treatment, or hospitalization diagnosis for inpatients, and assured and post-diagnosis for outpatients) prior to the onset of the ADE of interest. Administered drugs were obtained from reimbursable dispensation data prior to the onset of the ADE of interest.
- *Data acquisition and transformation:* In general, data acquisition and transformation were designed such that entire tables and relevant fields could be extracted using the database connector `ROracle`. Then the retrieved data were cleaned including variables and dates recoding, and joined in `R` (v3.4.3). First, GePaRD was queried for patient confounding variables, insurance periods and death dates. A temporary table with eligible subjects was saved to the project schema in GePaRD to speed up joining using `SQL` and reduce the operation time and load of `R`. Second, GePaRD was queried for drug dispensation for the eligible subjects. Dispensations between 01-01-2015 and 31-12-2016 were translated into ATC code using data from the in-house drug reference database (tab-delimited flat file). Third, GePaRD was queried for in- and outpatient diagnoses for the eligible subjects. Diagnoses between 01-07-2014 and 31-12-2016 were subsetted to cover both the baseline and the study periods. To overcome OOM runtime error due to data dimensions, only for outpatient diagnosis, data subsetting and cleaning were performed using `SQL` queries. Fourth, diagnoses data were merged, cases and eligible controls were identified, and index dates were retrieved for cases. Fifth, training and test data sets were created from cases that were matched with controls by parallel filtration of patients' birth year and sex using `doParallel` package.

Only socio-demographic data were used for filtration to reduce memory load. Objects containing eligible subjects diagnoses and dispensations were temporarily stored as single `R` objects in `.Rds` files on the computing cluster prior to transformation.

- *Data preparation for analytics:* First incidence of diagnosis and dispensations until the respective index date of the case and control were selected. Data on covariates were split into chunks and transformed into binary indicator (i.e., boolean) matrices that were reduced to one matrix for each of the training and test data sets. Zero variance predictors (i.e., all 0 or all 1) were excluded from the training and test data sets. Dispensation and diagnosis data were treated either independently or aggregated according to TTD human target pathways. Single-member groups were created for singletons including birth year, sex, and covariates that had no pathway-based group for group/block-based methods group lasso, block forests and ARTP. Single-member groups increased the number of groups by almost 10-fold (see Subsection 4.2.3).

3. **Functional target data:** TTD (update: 6.1.01; published: 2017.10.04) data were downloaded and loaded into a portable light relational database (`SQLite`) with `Bash` scripts. An `R` program was developed that connects to the `SQLite` database to query the data as necessary. Data were extracted, cross-matched with ATC, ICD and KEGG pathway ID, and transformed into binary matrices and key-value `data.table` structures of drug-disease, drug-target and disease-target pairs. TTD classification neither provides a categorization of all ICDs (or ATCs) nor categorizes ICDs at equal levels of hierarchy. ICDs in TTD groups (ICD-10-CM version 2017) were inflated to and cross-referenced with ICD-10-GM version 2017 down to the lowest level in the ICD hierarchy. Both the database and the matrices are contained into the `R` package. Figure 4.4 is a simplified entity-relationship diagram illustrating the `SQLite` database structure.

4. **Statistical modeling and prediction:** As mentioned in the previous subsection, a number of statistical tests were considered for this study to compare classical PV methods against GSEA-based methods and to investigate the effect of grouping the covariates. In general, wrapper functions were developed to 1) read training and test data sets, 2) assign the covariates to TTD groups, 3) use the training data set for fitting the models, 4) use the test set for predicting the outcome, 5) calculate the recall, precision and F1-score for each method, and 6) extract, if applicable, the most informative groups/variables. Here, I list the methods, their parameters and specific considerations regarding data preparation if necessary.

    - *Standard model for logistic regression:* Here, a logistic regression model was

applied using `stats` to analyze the relationship between each covariate and the ADE; p-values were adjusted for multiple comparisons according to Bonferroni (cutoff < 0.05).

- *The lasso:* The implementation in the `R` package `glmnet` (v2.0-16) was used with parallelization option, employing a 10-fold cross-validation to select the tuning parameter, λ, that minimizes the deviance.
- *The lasso for a constructed group variable (NGL):* The group variable was calculated as the sum of the covariates within a group and multiplied by 1 / square root of the group size. The calculation of the group variable was parallelized using `parallel`.
- *The group lasso for overlapping groups (OGL):* The regularized regression methods group (Breheny and Huang, 2009; Friedman et al., 2010) and overlapping group (Zeng and Breheny, 2016) lasso, among others, provide additional regularizations on group membership by using different penalty functions. We tested `grepregOverlap` (Zeng and Breheny, 2016) (v2.2.0) in `R` on simulated data (Appendix A.2).

  `grepregOverlap` for overlapping group lasso uses `grepreg` (Breheny and Huang, 2009), yet, to handle overlapping groups, the input design matrix is inflated (i.e., expanded) prior to the fitting and the prediction. It was not possible to use the current implementation of `grepregOverlap` due to the input and output data size (as data.frame) and the change in data type and/or object class within the `grpregOverlap`. Therefore, I adapted the matrix expansion function to handle sparse matrix objects as both input and output. However, the resulting matrix dimensions were not computable using `grpreg` (v3.2-1). I, therefore, used `pyglmnet` (based on (Friedman et al., 2010)) in `Python`, by adapting it to support sparse matrix input and intermediate calculations in order for the computation to be doable with the available amount of main memory (98 GB). A 10-fold cross-validation was employed to select the tuning parameter, λ, that minimizes the deviance. To the best of our knowledge, the modified `pyglmnet` is the only implementation with support for sparse matrices and cross-validation calculation of λ.
- *The pathway analysis by adaptive combination of rank truncated product (ARTP):* The ARTP is a gene set enrichment method that was originally designed for single nucleotide polymorphism (SNP) data (Yu et al., 2009). It is a hypothesis testing approach used to select the biological pathways that are enriched with genetic variants to be associated with a phenotype. The method preserves the correlation structure between genes by using permutation tests, and it has the potential to detect subtle effects of genetic

variants in a given pathway that might be missed when assessed individually. The ARTP uses p-values from any statistical association test, here a standard logistic regression model. We modified the ARTP to detect associations between ADEs and functional targets when using binary health care claims data. Here, cutoff for group selection is p-value ≤ 0.05, and number of permutations = 50.

Our previous implementation of ARTP (Appendix A.2) suffered from object class changes and growing objects within the function; both resulted in OOM runtime error in fitting. In addition, the implementation suffered from exceeded execution time in prediction. Three strategies were tested and implemented to adapt the ARTP to high-dimensional data, and reduce execution time and memory requirements: 1) minimization of memory fragmentation and growing objects by creation and initialization of all objects (e.g., data.frame, matrix and vector objects) within the fitting and prediction functions to preserve their size and dimension, 2) reduction of execution time by parallelization of permutation tests and prediction, and 3) utilization of on-disk storage of intermediate objects (e.g., predictions based on each group covariates) in case any forked process is killed, which is random to some extent.

- *Block forest (BF):* As a machine learning approach, block forests are a further development of random forests that is able to combine different blocks of omics data for outcome prediction and including group structures to improve the prediction performance. This is facilitated by modifying the split point selection procedure of random forests to the group structure in the data. BF handles the blocks independently, therefore, overlapping group structures can be analyzed without further modifications. However, singleton variables are not included in the analysis. The available implementation of BF could only be applied for risk prediction as it does not allow for variable or block importance estimation, and therefore, it is not used for variable and group selection.

  The `R` package `BlockForest` (v0.2.3) was used with number of tuning parameter sets = 50, number of tuning trees = 50, number of trees = 500, and splitting rule = 'gini'. BF performance, however optimized for parallelization, was hindered by memory and walltime restrictions. Adaptation for input as matrix was encouraged and implemented in the BF version to avoid memory cost of object class change between data.frame and matrix inputs throughout the operations; the newest version was used. Moreover, a parallelization of the tuning process was used to overcome exceeded execution time issues. The tuning parameters for the best run (minimum of all) were used to construct the forest.

4.2.3 Evaluation

Secondary data, including EHDs, are a rich data source that can be used to answer multiple research questions, using various analysis and mining methods. This allows for knowledge and method transfer between fields, which often requires various adaptations to the methods. In this case study, various adaptations were applied and their outcomes were cumulatively evaluated in regard to 1) data acquisition and management approaches with respect to data source, and 2) data dimensions and statistical methods performance.

Data source and dimensions drive acquisition and management approaches

Functional target data source: KEGG, STITCH and TTD: Developing the strategy for target-based enrichment analysis (Figure 4.3), KEGG, STITCH and TTD were considered. The functional target data source and the data acquisition and transformation approach largely influence each other. First, KEGG, the most comprehensive source for biological pathway data, offers various accessibility options. In addition to file-based (i.e., FTP) and graphic-based (i.e., XML representation of KEGG pathway maps) formats, KEGG databases can be queried in `R` for each drug-target (gene or pathway) and disease-target interaction using KEGG API, a `RESTful` web service application programming interface and the client-side package `KEGGREST`. Web service-based API insures dynamic data acquisition and is bandwidth efficient, however, it is network-dependent. Second, STITCH, a source for drug-drug similarity scores based on curated evidence-based drug-target interactions, could be downloaded as parsable flat files or as complete SQL schemas; the latter require large disk space and a robust free database management system (e.g., `MySQL`). Third is TTD, the source used in this chapter. A small number of flat files (e.g., STITCH) could be readily parsed in `R` or `Bash`, however, to better handle a complex structure such as TTD, an relational database management system is a better fit (Figure 4.4). Taken together, data source choice drives data acquisition and transformation approaches.

Functional target data management: Relational databases and MapReduce: To extract, transform and load TTD data, two possibilities were considered: relational database management systems (e.g., `ORACLE` or `SQLite`) and the well-known parallel framework MapReduce. The MapReduce programming model was developed and implemented for parallel and/or distributed processing of large-scale data in a `key/value` pair format (Dean and Ghemawat, 2008). Hadoop is the most popular publicly available implementation of MapReduce, which is based on the Hadoop distributed file system (HDFS) (Stonebraker et al., 2010; Muhammad et al., 2017). MapReduce is an ETL system that is often upstream from database management systems, and thus complementing them. MapReduce is best used in cases of ETL processes, complex data flows, semi-structured data, and the need for an out-of-the-box system for budget-limited projects (Stonebraker et al., 2010), which were not applicable to

our case study. `SQLite` has multiple features in favor of its use: 1) the database is a single file which is easily portable to many platforms without installation, 2) it is free and open-source (i.e., a good choice for low-budget small scientific projects), 3) it is simple to set up and query from most programming languages and environments, 4) it has the least possible dependencies for data manipulation, 5) it is straightforward to install and configure if required, and 6) it is simple with respect to data import, which is important for updating the functional target data with a newer version of TTD. `SQLite`, however, has limited scalability in case of multi-user concurrent workloads on the database; in these cases, `PostgreSQL` or `MySQL` could be more appropriate alternatives.

Longitudinal data management: High-dimensional data in `R` and `SQL`: To extract and transform GePaRD data into binary matrices, the inputs for the statistical methods, two possibilities were considered: `R` and `SQL`; in practice a combination of both was used. Using `SQL` bypassed memory and walltime constraints, however, resulted in poor scalability and reproducibility. To achieve scalability and reproducibility, as a practice, perhaps `R` could be used as a wrapper for `SQL` functions in future implementations. In addition, intermediate tables, on-disk storage of intermediate `R` objects and memory-mapped file objects could have been better utilized.

Data dimensions drive analytics choice and implementation

Case study statistics: The highlights of case study statistics are in Appendix A.3. Table A.3.3 shows the data dimensions of the eligible subjects. The socio-demographic as well as dimension statistics of the matched case-control data set are shown in Table A.3.4. The statistical methods performance metrics are in Table A.3.5. The total number of eligible subjects was 7,420,946; 1,159 (0.015%) died before the cohort exit date, 8,120 (0.11%) were not ADE-free after three months of cohort entry date and therefore excluded, while 11,732 (0.16%) suffered from the ADE afterwards. To assess whether our matched case-control design helped balancing patient time segments, patient time was calculated in calendar quarters for cases and controls and is presented in Figure A.3.10. The dimensions of the functional target groups (i.e., pathways) were assessed as well. The curated TTD data set contained 1,124 drugs and 12,761 diseases to the lowest level in 260 pathways; pathway sizes ranged between 3 and 10,511 (mean $\pm$ SD: 3114.1 ± 2463.1). Zero variance predictors were excluded (2,340 out of 12,407; 18.9%). A total of 10,064 informative covariates, excluding confounders, were mapped to TTD pathways; 8,161 (81%) were assigned to 260 pathways, while 1,903 covariates were not assigned to TTD pathways. Group sizes ranged between 2 and 6,441 (mean $\pm$ SD: 1908.5 ± 1468.9). If we accounted for the overlap among the groups, the actual dimensions of the 260 groups would be 496,210 covariates. The empirical distribution of the number of drugs and diseases in all functional target groups in the curated TTD data set was assessed and is presented in Figure A.3.11, and that in the GePaRD data set is presented in Figure A.3.12.

Data dimensions and analytics: The data in this case study influenced the analytics in two ways. First, the methods selection was data-driven. For instance, overlapping group logistic regression was considered instead of group logistic regression when preliminary research showed that most of biological targets overlap, i.e., drugs and diseases affect more than one pathway (e.g., drug ADEs). Second, the statistical methods were largely affected by data dimensions. Three factors influenced the data dimensions, and therefore affected the applicability of carrying out the statistical methods, the computation time and the runtime errors. First, the matched case-control study design contributed largely to reducing the number of subjects and number of covariables, and to balancing patient time. Second, the exclusion of zero variance predictors reduced the data dimensions and computation time, and enhanced the comparability of the tested methods; penalized regression-based methods ignored zero variance variables by default, while the other methods did not. Third, the qualitative assessments highlighted that the size and number of groups affected the statistical methods applicability and runtime. The number of groups increased by the introduction of single-member groups, which affected the ARTP and BF the most. It increased the number of subsets to be tested for association in ARTP, however, the smaller the group, the faster the standard logistic regression model ran. The number of groups increased the number of splits in BF and therefore the runtime. Group size affected the current implementations of group lasso the most; expanding the covariate matrix into a single much larger matrix perhaps is not the optimal solution for high-dimensional data. Benchmarking is required to further investigate the critical point at which the number of groups and covariates can no longer be analyzed by `grpreg`. Looking at the performance metrics, the worst methods with respect to prediction precision were the ARTP and OGL (Table A.3.5). Despite the adaptations for large-scale data and their performance with simulated data (Appendix A.2), those methods did not seem to have a fair chance against better implemented and large scale-adapted methods (e.g., BF). The ARTP would benefit from increasing the number of permutations, which would, however, increase memory and processing load.

4.2.4 Critical appraisal

There are crucial limitations in the implementation for this case study at three levels: longitudinal data management, the utilization of optimal parallelization options in `R`, and the evaluation. First, in a matched case-control study, each case is matched (i.e., paired) to a number of controls. In case no predictors were found between cohort entry and index date, a new match is assigned. In the current implementation, the predictors were not checked in the matching phase due to dimensionality, and therefore, memory issues. Thus, in case no predictive factors were available between the cohort entry date (01-01-2015) and the index date for either a case or a control subject, the pair was not regenerated. This resulted in an imbalance in

patient time segments. Instead of case:control 11,732:117,320, the final data set contained case:control 11,717:108,747, excluding 8,573 controls and 15 cases due to no predictors detected.

Second, there are several approaches for high-performance computing in `R`. The package `parallel` was used to improve the scalability of prediction by the ARTP. `batchtools` (Lang et al., 2017) was used for BF tuning parallelization. The promising approach by `future` (Bengtsson, 2019) was tested as a parallelization approach for ARTP prediction, however, its performance was completely halted in case any forked process was killed due to exceeded execution time or memory requirements. Therefore, possible approaches could be chunking the data into groups *a priori* to avoid memory load or using `batchtools`.

Finally, benchmarking was best to be used for evaluating the outcomes of this case study. This means: 1) evaluating the failed and successful implementations with respect to memory and time consumption, and 2) investigating the effect of number and size of groups and covariates on the implementations using simulated data.

4.3 Concluding remarks

Adverse drug events are a burden on the health care system that can be minimized through drug safety studies. The increasing availability of comprehensive secondary health data sources (i.e., electronic health care databases) and evidence-based genomic data represents a unique opportunity for advancing drug safety studies provided that methods for data integration and analytics are developed and adapted for scalability.

This case study attempted at applying and improving current data preparation and analytical methods for predicting the risk of ADEs on a large scale, incorporating knowledge on drug and disease molecular pathways. The case study highlights a number of consideration that are important to the success of large-scale drug safety studies. First, high-performance computing solutions are required for both data preparation and analytics. Data extraction and transformation for large-scale cohort studies can be memory consuming and require robust tools and better pipeline design. As well, statistical methods performance is limited by computational resources constraints, which requires scalable implementations. Second, signal detection in longitudinal data is driven by data source, structure and dimensions; it requires knowledge and method transfer as well as data integration from different fields. Third, at the analytics-level, this case study aimed at comparing the predictability of group-based (BF, ARTP, NGL, OGL) to that of the lasso and the standard logistic regression model. Despite their promising performance on simulated data, ARTP and OGL were hindered by data dimensionality. The group-based method, BF, outperformed all others in terms of sensitivity (i.e., recall), however, with respect to precision, the lasso and the standard logistic regression model showed best

performance. Indeed, evaluating the performance of novel methods requires more scalable implementations. In fact, drug safety studies would benefit from alternative high-performance computing platforms. Successful discipline-independent solutions were developed and applied to scientific data augmenting the stability and efficiency of relational database management systems with tools for large-scale data processing and analytics capabilities [e.g., SciDB (Stonebraker et al., 2011) and Array SQL (Misev and Baumann, 2015; for Standardization, 2019)].

This case study particularly shows that scalable implementation of statistical methods is the limiting factor in analyzing high-dimensional data. It also argues that a data-driven choice of acquisition and transformation tools would help expanding the scale at which drug safety studies are conducted (e.g., possibility to increase cohort size and integrate multiple data sources). Nevertheless, integration of data from multiple data sources can be as challenging; it requires models for integration to distinguish between informative and irrelevant attributes, and to achieve optimal knowledge extraction. In Chapter 5, I apply the principles of meaningful use of data to integrate multi-omics data from a cohort study. Having been subjected to large-scale studies in this chapter, the next chapter uses a pilot study to focus on the data integration and analytical challenges that are anticipated with the rise in multi-omics data availability.

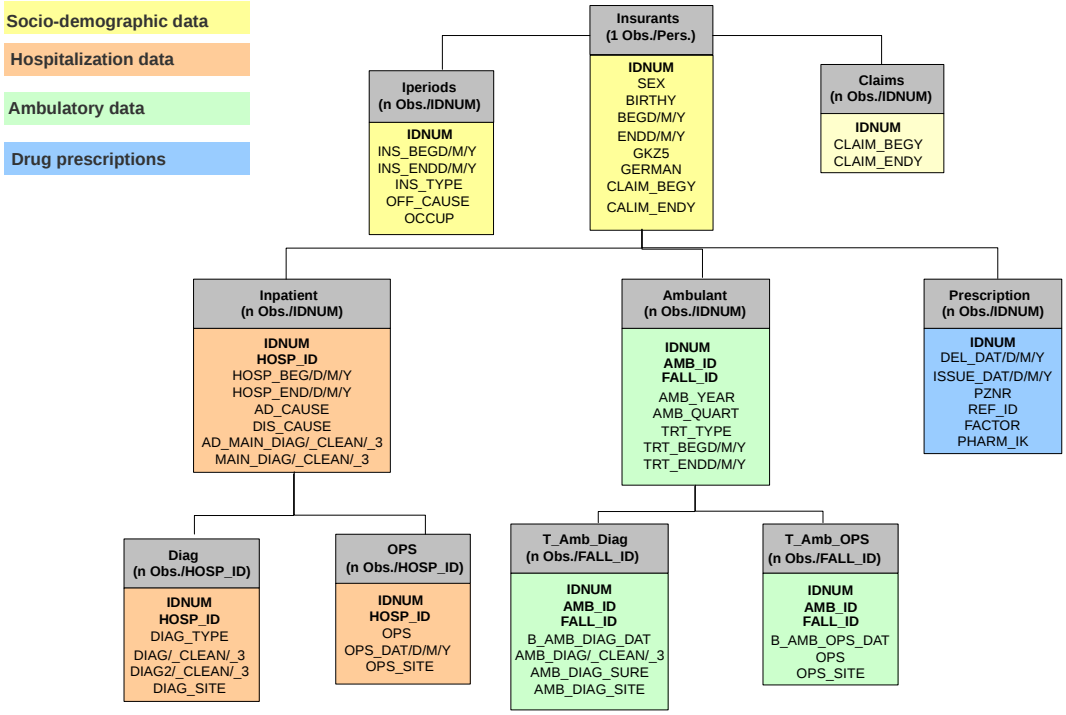
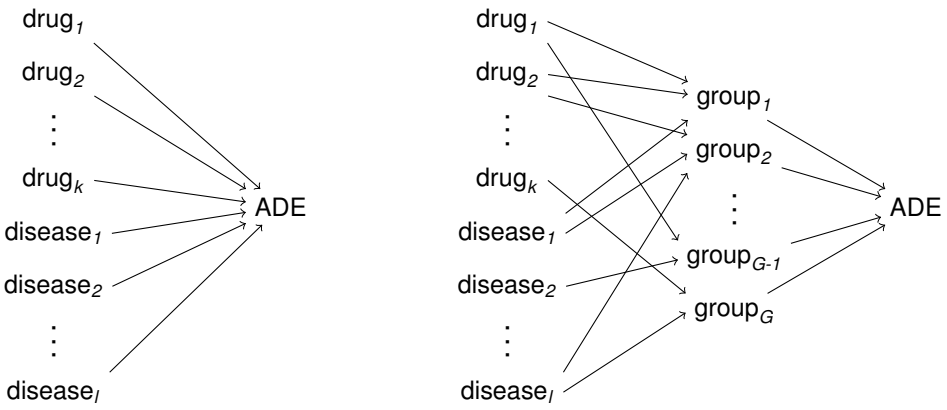


FIGURE 4.1: A simplified entity-relationship diagram representing the structure and relevant content of GePaRD. First, the socio-demographic dimension consists of demographic data on insurants (Table Insurants), insurance periods (Table Iperiods) and claims periods (Table Claims). Second, hospitalization (i.e., inpatient) data are linked through hospital ID and stored in: Table Inpatient (for hospital admission and discharge causes and dates), Table Diag (for diagnosis types and codes), and Table OPS (for procedures codes and dates). Third, ambulatory (i.e., outpatient) data are linked and stored in: Table Ambulant (for treatments), Table Amb_Diag (for outpatient diagnoses and dates), and Table Amb_OPS (for procedures). Finally, prescription data (central pharmaceutical reference number; PZN, and delivery and issue dates) are stored in Table Prescription, which can be linked to the dispensing pharmacy by pharmacy ID.



(a) without groups **(b)** with groups

FIGURE 4.2: A schematic representation of two approaches for ADE risk prediction. **(a)** illustrates the standard approach in the field of PV. The left column contains all the drugs $(1, 2, \ldots, k)$ and diseases $(1, 2, \ldots, l)$. The ADE of interest is shown on the right. The predictions are based on the associations between individual risk factors (drugs and diseases) and the ADE, represented here by arrows pointing from each drug/disease to the ADE. **(b)** illustrates the approach proposed in this case study. Similarly, the left column contains all the drugs $(1, 2, \ldots, k)$ and diseases $(1, 2, \ldots, l)$ as covariates. The middle column lists groups $(1, 2, \ldots, G)$. Each arrow between a drug/disease and a group represents the group membership. Note that drugs/diseases can belong to multiple groups simultaneously, e.g., drug_2 is in, both, group_1 and group_2. Instead of assessing the associations between the drugs/diseases and the ADE directly as in **(a)**, the associations between the groups and the ADE are assessed, shown here by arrows pointing from the groups to the ADE.

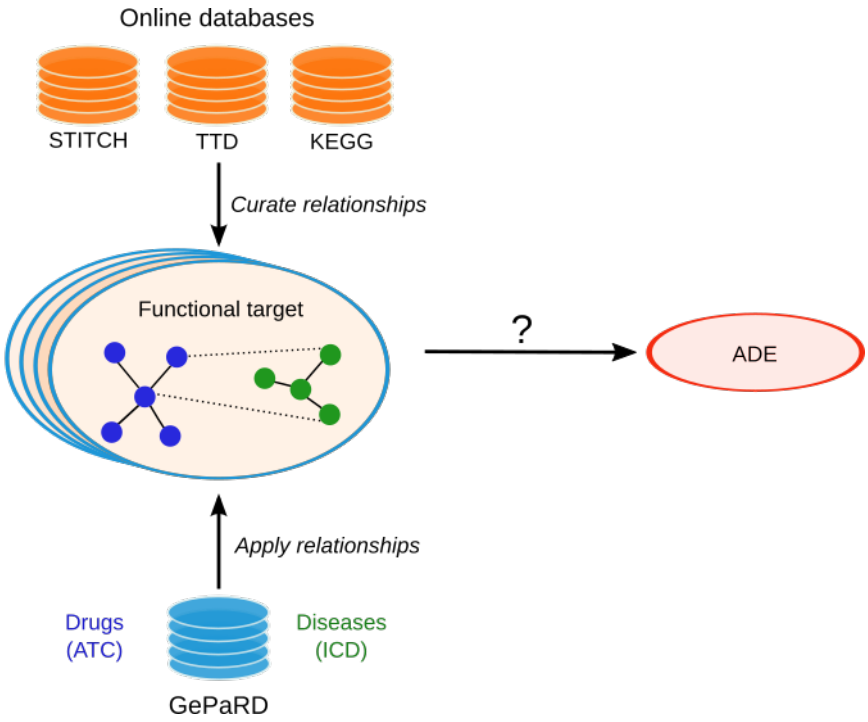


FIGURE 4.3: The procedure of an enrichment analysis to predict ADEs in routine data of the SHIs using functional targets (FTs). First, relevant online genomic knowledge bases are queried for drug-target, disease-target, drug-disease and drug-drug relationships to curate FTs. FTs serve as the grouping structure of the predictors. Within those FTs, substructures and pairings exist, such as drug-drug structural and functional similarity, drug-disease relationship, and less likely disease-disease co-existence. Second, an epidemiological study design is considered, and SHIs database (here GePaRD) is queried for prescribed drugs, in- and outpatient diagnoses that are coded according to international coding systems to facilitate being mapped to FTs. Third, the SHIs predictors are grouped according to the grouping structure, and the risk of ADE is predicted using statistical models based on those structures. Drugs are denoted in blue, while diseases are in green. Within a FT, solid lines represent drug-drug or disease-disease relationships; dotted lines represent drug-disease (i.e., indication) relationships. GePaRD = The German Pharmacoepidemiological Research Database, ADE = Adverse Drug Event, STITCH = Search Tool for Interactions of Chemicals, KEGG = Kyoto Encyclopedia of Genes and Genomes, TTD = Therapeutic Target Database, ATC = The Anatomical Therapeutic Chemical Classification System, ICD = The International Classification of Diseases. The figure is modified after (Foraita et al., 2018) (Appendix B.2).

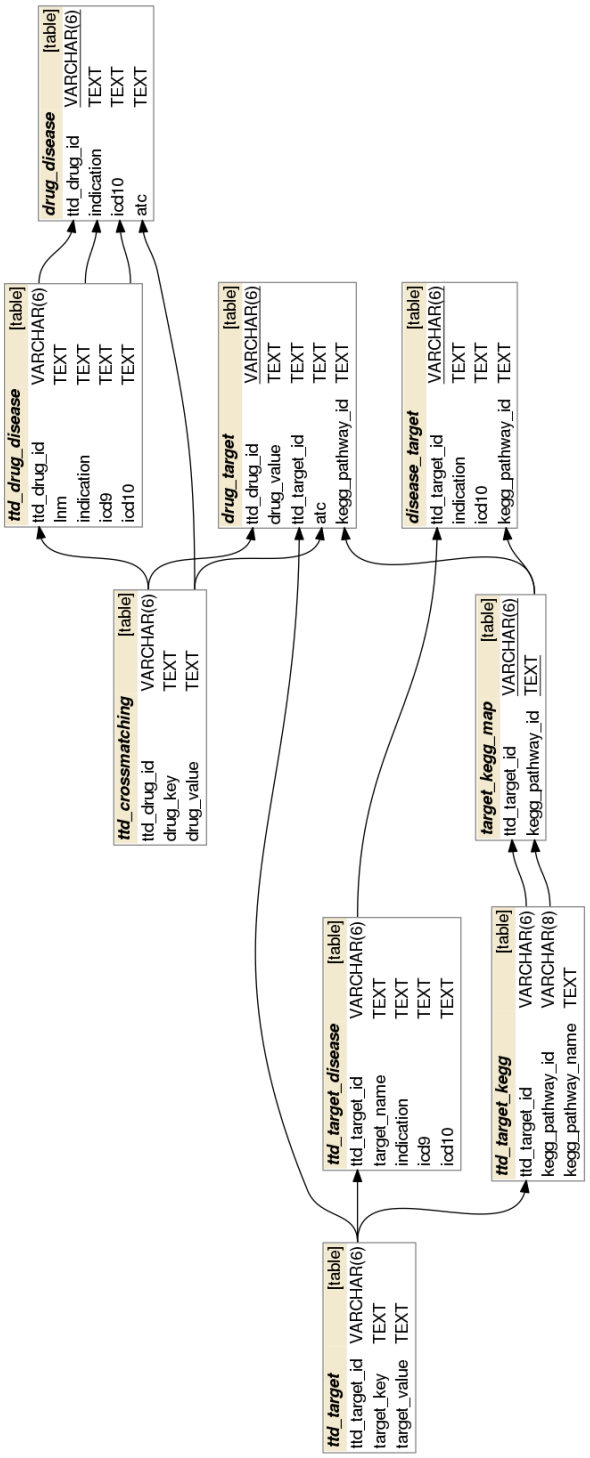


FIGURE 4.4: A simplified entity-relationship diagram (ERD) of TTD tables. Original data downloaded from TTD are used to create the tables starting with *ttd_*. The remainder of the tables are the result of data cleaning, integration and cross-referencing in `R` using `RSQLite` and `data.table` packages. The ERD was generated from the `SQLite` database using SchemaCrawler (v15.04.01). The data type of each column was pre-defined in *ttd_* tables using `SQLite`. In the tables created by `RSQLite`, the default is TEXT; the data type is manually corrected and denoted by the underlining in the ERD. The arrows represent the relationships between the tables based on entry identifiers.

Chapter 5

Meaningful Data Integration: A Case of Primary Health Data

Cohort studies are the heart of non-experimental epidemiological research. They offer insights into associations and causal relationships between lifestyle and, for instance, complex diseases and drug response in the population. This, in turn, advances our understanding of disease etiology and consequently drives preventive measures development, therapeutic decision making and drug development. The longitudinal design of such studies, where individuals serve as their own controls, facilitates correcting for intra-subject variability and investigating disease progression as well as its influencing social and behavioral factors in individuals in a time-dependent manner. In the omics era, large data volumes are generated from extensive multi-omics phenotypes of the individuals. Cohort studies combining omics data with other phenotypic information, such as lifestyle information, face challenges in the areas of data management, measurement, storage and analysis. Moreover, the omics era increases the need for additional resources for data acquisition, in particular sample collection and storage (e.g., in biobanks). It is, therefore, of utmost importance to adequately prepare for the multi-omics era of epidemiological research. Advances in high-throughput technologies rapidly accelerate the establishment and extension of longitudinal multi-omics biobanks around the world (Wijmenga and Zhernakova, 2018). Such resources increase the depth of the phenotypic profiles of individuals.

This chapter explores the different facets and data blocks in modern cohort studies using a case study based on the pan-European IDEFICS/I.Family cohort, where children have been extensively examined in a baseline and follow-up surveys. The study is an example of primary human data collected for answering particular research questions, namely on the etiology and primary prevention of childhood obesity and other metabolic disorders in Europe. The results of the case study will be prepared for submission to the journal *OMICS: A Journal of Integrative Biology*.

Using this case study, I investigate the applicability of dimensionality reduction

approaches in achieving meaningful use and integration of heterogeneous high-dimensional data, which is expected to be a typical case in future epidemiological studies. The chapter also highlights self/past-dependency issues in data acquisition practices from shared data repositories, transformation and integration of heterogeneous multi-omics data and classical epidemiological data, statistical modeling, and association and interaction analyses.

The chapter is structured as follows: Section 5.1 gives background information on multi-omics biobanks, cohort studies and plasma lipidomics data in epidemiology, and a brief introduction to the IDEFICS/I.Family cohort and the MyNewGut project. Key aspects and applicability of meaningful data integration are also discussed. Section 5.2 addresses the chapter objective in terms of project requirements, solution implementation and results evaluation. Concluding remarks are presented in the last section.

5.1 Background

5.1.1 The rise of multi-omics biobanks in cohort studies

Biobanks are defined as the "structured resources that can be used for the purpose of genetic research, including human biological materials and/or information generated from genetic analysis and associated information" (Hewitt and Watson, 2013; Coppola et al., 2019). Biobanks facilitate understanding the etiology of complex diseases, advancing personalized medicine research and driving drug development. Taking a closer look into biobank components, biospecimens collected and stored from large-scale cohorts include tissues, saliva, urine, stool, and blood. From those biospecimens, DNA, RNA, metabolites and proteins can be extracted and as well stored. Software programs were developed and used for sample management, retrieval and transfer (Coppola et al., 2019). Biobank data management, however, seems to be a challenge in respect to: 1) research data management, 2) real-time data sharing and 3) disaster management. Cloud-based solutions are suggested to overcome issues of scalability and disaster management (Paul et al., 2017). Moreover, resources for harmonization of biobank data and operating procedures are reviewed in (Harris et al., 2012).

According to Coppola *et al.*, biobanks can be classified based on: 1) design (e.g., population, disease-oriented), 2) purpose (e.g., epidemiological or pharmaceutical research), and 3) study type (e.g., family cohort studies, clinical trials) (Coppola et al., 2019). Population-based biobank samples are collected from volunteers, as it is the case in the German National Cohort (German National Cohort (GNC) Consortium, 2014). Population studies can also exploit the family aspect as it is the case in the pan-European IDEFICS/I.Family study (Ahrens et al., 2017).

The vast majority of biobank samples and data originate from cohort studies. Driven by the advances in genotyping and genome sequencing, cohort studies have increasingly adopted multi-omic approaches (Huang et al., 2017; Hasin et al., 2017; Wijmenga and Zhernakova, 2018) to shed light on the different facets of the biological system such as genetics, epigenetic modifications of DNA, functioning molecules (peptide, lipid and metabolite abundances), and microbiome composition and function (Hasin et al., 2017). The integration of these facets deepens our understanding of the individuals and of their behavior in relation to complex diseases.

Although beneficial, using omics in cohort studies can have its challenges. First, in spite of the reduction in cost of multi-omics laboratory analyses, cost is still an obstacle, in particular, if a large number of participants is required (such as in predictive and biomarker discovery studies). Second, cohorts can be less diverse and therefore show less variability in geno- and phenotypes (as it is the case in small closed countries). Third, cohort studies have particular aims that shape the cohort study protocol (e.g., age, sex, number of participants and sets of phenotypes); therefore it is desirable to combine smaller specific cohorts into a single larger cohort to answer biological questions. At that point, however, data harmonization between these smaller cohorts becomes an issue (Wijmenga and Zhernakova, 2018). Forth, omics data are heterogeneous in nature, and each omics data type presents challenges in data analytics, in particular in selecting the variables associated with the outcome, and in differentiating between causal relations and associations (Hasin et al., 2017). Such heterogeneous omics data, possibly also from heterogeneous cohorts, are required to be integrated and included in statistical models for an "integrative holistic" omics approach for analysis (Hasin et al., 2017).

5.1.2 Data heterogeneity and meaningful use of data

Several characteristics define heterogeneous data; those include high variability, ambiguity, large fraction of missing values, and high redundancy (Wang, 2017). Reasons for data heterogeneity can include the diversity of data acquisition devices. There are various types of data heterogeneity, including the following relevant types: syntactic (different languages for different sources) and terminological (different names for the entities from different data sources) (Wang, 2017). Cohort omics data suffer from both types of data heterogeneity. In handling heterogeneous data, three levels of processing are considered: 1) cleaning, 2) integration and 3) dimensionality reduction. First, at the data cleaning level, utilized methods encompass de-duplication, imputation of missing values and correlation analyses to distinguish between informative and irrelevant attributes. Second, at the data integration (i.e., aggregation) level, data sets are matched and merged to provide a data set that can be used for data mining. Third, dimensionality reduction and data normalization techniques are utilized to minimize the computational burden and to find meaningful informative patterns in the data (Wang, 2017).

However important, going a step further beyond cleaning and integration must be considered. Without the implementation of "meaningful use of data" (Bizer et al., 2012) principles, the application of aforementioned data engineering approaches yields sub-informative data. To promote meaningful data integration, according to Bizer *et al.*, several steps can be undertaken including: 1) problem definition, 2) database query for candidate data elements required to investigate the problem (e.g., search the database for all patients taking a certain drug), 3) implementation of ETL workflows to transform the relevant data into an appropriate functional format, 4) entity resolution including data verification and abstraction, and 5) implementation of appropriate statistical and computational methods for problem solving (Bizer et al., 2012).

Modern cohort studies often require the integration of various data types that are extremely heterogeneous, for instance, from multi-omics studies, lifestyle variables from epidemiological profiling and surveys, exercise data from wearables, and food intake data from food surveys and food tracking web applications. All is required to be integrated into a meaningful data model, which is of utmost importance to achieve optimal knowledge extraction. Data engineering offers principles for the meaningful use of biological data.

5.1.3 Lipidomics in epidemiological research

Lipidomics is defined as "the characterization, analysis and study of the lipid complement of biological systems" (e.g., tissues or fluids) (Mundra et al., 2016). Lipids and fatty acids are crucial substrates to humans. They are involved in energy production, biological membranes construction, and signaling molecules. Lipids also serve as therapeutic drug targets. Estimates of the number of lipid species in nature range between 10,000 and 100,000 (Wenk, 2010). Lipids are not encoded in the genome, they are molecules that rather result from metabolic processes (Wenk, 2010), for instance, the metabolism of dietary fat by the digestive enzymes or by the gut microbiota (Wolters et al., 2019).

Mass spectroscopic analysis of lipids was first used in the 1990s. Following that, mass spectroscopy-based lipidomics techniques were developed (Wenk, 2010). High-resolution mass spectrometry allows for quantification of lipids, discrimination between lipids with similar masses and chemical structures, and identification of novel uncharacterized lipids (Wenk, 2010). Data analytics in lipidomics comprise data processing (mass-spectroscopic peak identification and normalization), statistical analysis (Datta and Mertens, 2017), and elucidation of biological relevance via integration into known biological pathways and processes (Wenk, 2010).

Dissecting a lipidome, the main fraction of the plasma lipidome consists of lipoproteins: very low-density lipoprotein (VLDL), low-density lipoprotein (LDL) and high-density lipoprotein (HDL). These lipoproteins consist of the lipid classes: phospholipid, sphingolipid and free cholesterol, cholesteryl ester and triacylglycerol; those classes, in turn, consist of numerous lipid species. Plasma lipoproteins function as transporters of lipids between the gut, the liver and the peripheral tissues, therefore plasma lipidome analysis has a unique position in inferring relationships between lifestyle and genetic factors and metabolic processes. Methods for statistical analysis of lipidomics data in epidemiology are reviewed in (Mundra et al., 2016; Datta and Mertens, 2017).

5.1.4 The IDEFICS/I.Family cohort study

IDEFICS/I.Family is a pan-European population-based children cohort representing diverse European lifestyles that certainly affect lipidome and microbial diversity in children. The IDEFICS study aimed at understanding the etiology of childhood obesity in Europe, and examining the feasibility and effectiveness of primary intervention strategies concerning diet, sleep and physical activity in eight European countries: Belgium, Cyprus, Estonia, Germany, Hungary, Italy, Spain and Sweden. In the IDEFICS cohort, children were extensively profiled and examined in a baseline survey (T0; 2007-08), and a first follow-up examination (T1; 2010-11). A second follow-up examination (I.Family; T3) took place in 2013-14, where not only IDEFICS children were included but also their parents and siblings to investigate social and familial effects, as the I.Family study aimed at identifying determinants of lifestyle behaviors, in particular dietary behavior. The baseline survey included 16,229 children aged 2-9.9 years. At T1, 11,041 (68% of baseline survey participants) children and 2,555 newly recruited children were examined. At T3, 9,617 children and 7,941 adults were included in the follow-up examinations. A simplified illustration of the longitudinal design of the IDEFICS/I.Family cohort study is shown in Figure 5.1.

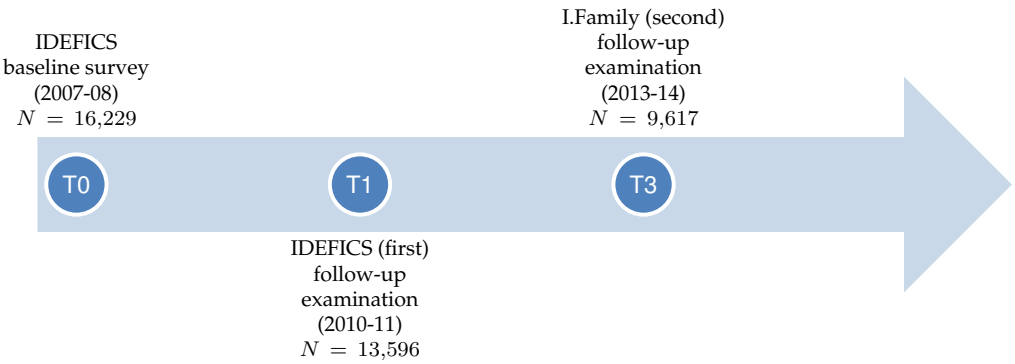


FIGURE 5.1: A simplified illustration of the longitudinal design of the IDEFICS/I.Family children cohort. It depicts the time points of surveys and examinations from which the data in this case study are derived. The figure is based on Figure 1 in (Ahrens et al., 2017).

The examination modules are described in detail in Ahrens *et al.* (Ahrens et al., 2017). The most relevant modules include: questionnaires (answered by parents or later by the participants themselves), physical examinations (anthropometry, blood pressure), the collection of non-invasive biomaterial (saliva for DNA analyses, stool for lipids and microbiota analyses, urine) and invasive biomaterial (fasting blood for biomarker measurements including lipidomics), as well as accelerometry (physical activity trackers). Questionnaires aimed at collecting information on: dietary behavior (food frequency and complemented by structured computer-based 24-h dietary recalls), physical activity type and duration (including sports club membership), sedentary behavior (media devices in child's room and screen time), and others (such as sleeping habits, medications and medial history, socio-demographic characteristics of the parents). Moreover, in the framework of the MyNewGut project, stool samples from IDEFICS/I.Family participants, among others, were sequenced and analyzed to identify microbiome-related features that contribute to and predict obesity and other disorders, and to understand the effect of environmental factors on gut microbial communities and its consequences on health outcomes (Sanz et al., 2018). However, for such a large cohort, resource allocation for sample collection, storage, processing, sequencing and chemical analysis form a bottleneck. Therefore, a pilot study was designed to analyze microbiome and plasma lipid profiles of a relatively small number of participants, and to investigate the potential of integrating lipidome and microbiome data.

5.1.5 Case study: Childhood obesity and associated markers in plasma lipidome and microbiome profiles

Childhood obesity has become an epidemic worldwide. In 2016, more than 340 million children and adolescents were overweight or obese (World Health Organization, 2021a). In addition to the general causes of obesity, namely increased fat/sugar/energy intake and decreased physical activity, a number of factors contribute to obesity in childhood and adolescence. These factors include: 1) prenatal factors (e.g., smoking during pregnancy, maternal diabetes, maternal excessive or reduced energy intake), 2) diet, familial, and in adolescents, social and environmental habits (e.g., irregular meal patterns) (Magrone and Jirillo, 2015; Ahrens et al., 2017), 3) genetics (Cugino et al., 2013; Iacomino et al., 2016), 4) gut microbiome as a key player in glucose and fat metabolism, and consequently in metabolic homeostasis (Rampelli et al., 2018; Wolters et al., 2019). Overweight and obesity in childhood is a risk factor for serious health outcomes in adolescence (Börnhorst et al., 2019) and adulthood (Magrone and Jirillo, 2015) including: insulin resistance, metabolic syndrome, type 2 diabetes mellitus, and cardiovascular diseases (Bremer et al., 2012).

Obesity is considered not only a metabolic disorder but also an inflammatory disorder as reviewed by Magrone and Jirillo (Magrone and Jirillo, 2015). A number

of inflammation-related markers were elevated in obese animals and humans, respectively, the adipose tissue-derived tumor necrosis factor (TNF)-α [reviewed in (Magrone and Jirillo, 2015)], and C-reactive protein [CRP; in children (Nappo et al., 2013) and adults (Vargas et al., 2016)] and the adipose tissue-derived interleukin (IL)-6 (Eder et al., 2009; Pradhan et al., 2001). IL-6 plays a key role in the biosynthesis of CRP. Moreover, obesity has been linked to immunological diseases such as asthma. IL-6 and IL-8 levels were elevated in obese asthmatic and non-asthmatic children compared with asthmatic non-obese children and control children (Magrone and Jirillo, 2015).

A number of molecular lipid species were demonstrated to be associated with weight status in adults. Recent lipidomic studies have shown associations of specific lipid species (cholesteryl ester, ceramide and lactosylceramide) (Cheng et al., 2015) or lipid classes (lower levels of glycerolipids but higher levels of glycerophospholipid) (Jové et al., 2014) with weight status in adults. Phosphocholine PC16:0/2:0 was negatively and PC14:1/0:0 was positively associated with visceral fat (Syme et al., 2016). As well, ceramides were associated with inflammation and insulin resistance (De Mello et al., 2009). Specific lipid classes were also associated with asthma. In particular, reduced levels of phosphatidylglycerol, ceramide-phosphates and ceramides, and increased levels of sphingomyelin 34:1 were found in the airway lipid particles in adult asthmatic patients in comparison to healthy adults (Hough et al., 2018).

The relationship between lipid metabolism and gut microbiota is very tight. Gut microbiota play a key role in lipid metabolism and energy homeostasis (Wolters et al., 2019), and lifestyle factors (diet, physical activity and sedentary behavior) influence both lipid metabolism and gut microbiota (Bressa et al., 2017; Wolters et al., 2019; Rampelli et al., 2018). Gut microbiota was shown to modulate lipid metabolism in mice (Velagapudi et al., 2010; Kindt et al., 2018). Markers in gut microbial genera were demonstrated to be associated with weight status in children (Rampelli et al., 2018). Imbalances in gut microbiota were associated with immunological diseases such as respiratory (allergic rhinitis and asthma) or dermatological (atopic dermatitis and eczema) allergies in infants (Chua et al., 2018), and food sensitivities in children (Savage et al., 2018). Studies investigating the interaction between gut microbiota and plasma lipids in children and their impact on weight status and immunological health are, however, scarce.

5.1.6 Study objectives

Biological objective: The present case study aims at understanding the associations of lifestyle factors (namely diet, physical activity and sedentary behavior) and plasma lipidomics with weight status and immunological health in children, and whether these associations are mediated by intestinal microbiota. This study serves as a pilot study; it is expected to increase our understanding of the aforementioned associations, potentially leading to improved lifestyle recommendations for children

and adolescents. The analyses are based on data that have been collected in the framework of the IDEFICS/I.Family children cohort (Ahrens et al., 2017) and analyzed in the framework of the European project MyNewGut (Sanz et al., 2018).

Data engineering objective: I use this case study to investigate the applicability of dimensionality reduction approaches in achieving meaningful use and integration of heterogeneous high-dimensional biological data. For this purpose, I employ dimensionality reduction approaches to analyze data from a number of heterogeneous sources (namely plasma lipidome, microbial abundances and epidemiological profiles). On this reduced data set, I apply a statistical model to infer associations between variables from multiple sources and the outcome (weight status and immunological health status). Throughout the chapter, I follow the steps discussed by Bizer *et al.* for meaningful data integration.

5.2 Meaningful data integration

As mentioned earlier, cohort studies require the integration of heterogeneous data into a meaningful data model for outcome prediction, effect estimation and statistical inference. Data engineering principles of both heterogeneous data handling and meaningful use of data can be applied in this context. To investigate the applicability of these principles, I chose the pilot study. Being a pilot study ($n = 70$), and due to the depth of the information on the pheno- and genotypes of the participants of the IDEFICS/I.Family cohort, the data are severely challenged by dimensionality issues.

This section addresses: 1) the challenges and requirements of this pilot study explaining the data sources to be integrated [i.e., part of problem definition as in (Bizer et al., 2012)], 2) the solution implementation steps in the light of data heterogeneity and meaningful use of data principles, 3) the evaluation of data preparation and analysis, and 4) the limitations of this case study and those of the solution implementation.

5.2.1 Challenges and project requirements

The biological aim of the study, as mentioned above, is to understand the associations of three integral lifestyle factors and plasma lipidomics with weight status and immunological health status in children, and to determine whether these associations are mediated by the diversity of intestinal microbiota. These associations are to be inferred using phenotypic information at two time points [T1 (2010-11) and T3 (2013-14)] on a group of children. The children were non-obese at T1 and almost half of them developed obesity by T3 while the other half maintained a non-obese weight status.

There are three aspects of the participants' phenotypes of interest here, namely: epidemiological profiles[1], plasma lipidome profiles and microbiome profiles. Each of these profiles come into play at a different step of the statistical analysis plan.

1. **Epidemiological profiles:** The IDEFICS/I.Family cohort is the main source for data on lifestyle variables and confounding variables (i.e., age, sex). Data retrieval, curation and preparation steps were necessary to use and compare data from two follow-up examinations. Table 5.1 shows the different variables required and the respective IDEFICS/I.Family mean of assessment.

   The cohort data are stored on a central data server hosted at BIPS. A time- and study time point-limited access to cohort data is granted after an evaluation of a project proposal. IDEFICS data (including T1) are stored as SAS7BDAT binary database storage files. I.Family data (including T3) are stored in a relational database management system (MySQL). The database systems contain, among others, sets or tables of metadata on participation and biological samples, and data from questionnaires, physical examinations, accelerometry, and 24-h dietary recall (24-HDR). Each participant is assigned a unique identification number (ID); the aforementioned tables or sets can be linked via this unique ID. Physical examination assessments [e.g., CRP (Schlenz et al., 2014)] were standardized according to age and sex, and tested for quality by BIPS before being stored in a designated MySQL table. Body mass index (BMI) was calculated from measured height and weight, and categorized according to (Cole and Lobstein, 2012).

2. **Plasma lipid profiles:** Results of a targeted plasma lipidome analysis of the participants represent a plasma lipidome profile for each participant at each time point. The lipidome analysis measured 328 lipid species (including 12 internal standards) present in eight classes ($n = 53$ at T1 and 55 at T3; $n = 45$ paired data points at the intersection of T1 and T3). The fractions of mono- and polyunsaturated and saturated fatty acids in each class are assessed as well. The data set is stored as Microsoft Excel tables.

3. **Microbiome profiles:** Results of the intestinal microbiome analysis of the participants ($n = 70$), published in (Rampelli et al., 2018), are used to assess the microbiome diversity at each time point. Proxies for microbiome diversity were to be calculated from microbial abundance tables at the genus-level (167 genera) stored as flat files.

Aspects of heterogeneity In this case study, heterogeneity is expected at two levels: syntactic and terminological. Syntactic heterogeneity stems from the different data storage systems used in the study, namely MySQL, SAS7BDAT, Microsoft Excel

[1] In the chapter, I use the term "epidemiological profile" to combine: 1) lifestyle variables (diet, physical activity and sedentary behavior variables), 2) demographic (age and sex), and 3) non-modifiable (i.e., social) variables (maternal BMI, puberty status and socio-economic status).

TABLE 5.1: Epidemiological profile components used in the case study and their respective IDEFICS/I.Family method of assessment.

Variable	Assessment
Socio-demographic variables	
Sex	Questionnaire
Age [in years]	Questionnaire
Country	Questionnaire
International Standard Classification of Education	Questionnaire
Puberty status	Questionnaire
Maternal BMI	Questionnaire
Clinical variables	
CRP z-score	Physical examination
BMI z-score	Physical examination
Physical activity	
Moderate-to-vigorous physical activity [minutes per day]	Accelerometry
Sports club membership	Questionnaire
Sports club time [minutes per week]	Questionnaire
Sedentary behavior	
Number of media devices in bedroom	Questionnaire
Screen time [hours per week]	Questionnaire
Diet	
Fish frequency [times per week]	Questionnaire
Preserved food frequency [times per week]	Questionnaire
Sweet propensity score (including diet soft drinks)	Questionnaire
Fat propensity score	Questionnaire
Fiber intake (g/day)	24-HDR
Usual weight of food intake (g/day)	24-HDR

and flat file. Even though IDEFICS/I.Family data passed rigorous quality control, standardization and data cleaning procedures, slight terminological heterogeneity is observed. Such heterogeneity stems mainly from the differences between T1 and T3 in variable names and variable content (e.g., difference in food groups between European countries, and improvements in physical examination procedures). It was of utmost importance to understand the levels of heterogeneity and carefully address them at the programming and knowledge extraction levels.

Statistical analyses A model for hypothesis testing is to be carefully chosen, and applied to the data to test for association of T3 data points (cross-sectional analysis), and the difference between T3 and T1 (longitudinal analysis). For that, each of the above listed data sources had to be retrieved, cleaned, transformed and reduced before applying appropriate statistical models for two continuous outcome variables: BMI z-score and CRP z-score. Moreover, data are to be segmented and described by weight gain categories (normal and overweight/obese BMI z-score at T3) and

immunological health categories (normal and high CRP z-score at T3; ≤ 0 and > 0).

Existing solutions R is a strong platform for both data retrieval and curation, and statistical programming. R can also handle such a diverse ecosystem of data sources effectively. Moreover, even though multivariate analysis solutions for lipidomics are available commercially, alternative open source and free solutions have recently become available in R. Therefore, an R package had to be developed to contain all data management, visualization and analysis steps to ensure reproducibility.

5.2.2 Solution implementation

The solution is designed to test the applicability of meaningful use of heterogeneous data through dimensionality reduction. The solution implementation is described according to (Bizer et al., 2012). Data preparation tools, data dimensionality reduction and analysis methods are highlighted here as well.

1. **Problem definition:** The main question, as described above, concerns the investigation of associations of plasma lipidomics with weight status and immunological health status in children, allowing for comparing those who developed obesity between T1 and T3 to those who maintained a normal weight. These associations can be mediated by the diversity of intestinal microbiota, and are sensitive to a number of lifestyle variables.

2. **Database query:** All children whose microbiome and plasma lipidome have been profiled were included in this study. Therefore, IDEFICS/I.Family data sets were searched for the IDs of those participants for data extraction.

3. **Data transformation:** A simple workflow was designed and implemented in R to dynamically:

   (a) retrieve respective relevant epidemiological data from IDEFICS/I.Family data sets using MySQL connectors and SAS7BDAT readers for R, and calculate variables from source (e.g., maternal BMI),

   (b) parse Microsoft Excel tables to extract lipid profiles at two levels: summarized lipidome variables by lipid class (i.e., saturated, mono- and poly-unsaturated fatty acid fractions of the lipid classes) and individual lipid species, both as percentages of total lipid class, and clean the data from internal standards, blanks and controls, and

   (c) parse microbial abundance flat file to extract microbiome profiles.

   The workflow was supplemented by a configuration file for database authentication information, and input and output data sets locations, intended at insuring data security and preserving the directory structure. Moreover, to promote dynamic and reproducible programming, a tab-delimited flat file was used containing the variable names and respective data set name or location in the

IDEFICS/I.Family data sets. Both files were parsed at runtime. The workflow is sketched in Figure 5.2.

4. **Entity resolution:** Bizer *et al.* considered entity resolution as the forth step in the meaningful use of data (Bizer et al., 2012). Entity resolution involves the extraction, matching and resolution of entities in data sources (Getoor and Machanavajjhala, 2012). It covers aspects of de-duplication, record linkage, verification of elements of each unique entity, and classification (i.e., canonicalization) by analyzing those elements across the different data sources at many levels of abstraction and from different perspectives (Bizer et al., 2012; Getoor and Machanavajjhala, 2012). Entity resolution is a challenge in many fields such as database management, machine learning, natural language processing, and statistics (Getoor and Machanavajjhala, 2012). Impaired entity resolution results in impaired knowledge extraction (Bhattacharya and Getoor, 2007). Given the small number of records ($n = 70$) and the high quality of the data sources, verification of the relevance and comprehensiveness of data as well as abstraction were most relevant to this case study. Below are the different steps in regard to entity resolution that are used to handle the three data sources.

   (a) The elements (i.e., variables) of the epidemiological profile were selected using domain knowledge to insure comprehensiveness and avoid redundancy. Relevant elements were verified using descriptive analytics across different weight gain and CRP z-score categories. For the abstraction of the epidemiological profiles, we aimed at constructing a lifestyle variable to describe each participant (i.e., to assign each participant a lifestyle category or class based on his/her observed socio-demographic, PA, diet, and SB observations). First, missing lifestyle data were imputed by predictive mean matching as implemented in the `R` package `missRanger` (v2.1.0) (Stekhoven and Buhlmann, 2012). Second, the data were scaled, and the optimal number of clusters ($k = 2$) was inferred using `NbClust` (Charrad et al., 2014) (v3.0) with the parameters `method = "centroid"`, `index = "alllong"`. A data-driven clustering, based on self-organizing maps (SOM), of the most informative variables (verified by Pearson test for correlation; namely fat and sweet propensity scores, maternal BMI, exercise duration, and number of media devices in bedroom) was adopted as implemented in `kohonen` (v3.0.8) (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018). Third, the SOM nodes were clustered into two clusters according to partitioning around medoids (i.e., PAM) as implemented in `cluster` (v2.1.0). A composite ID was created combining IDEFICS IDs and I.Family IDs to track the subjects across the clusters at T1 and T3 (i.e., whether the subjects will remain in a defined cluster).

   (b) The lipid profiles were abstracted as the summarized lipidome variables

of the different lipid classes (i.e., the eight lipid classes are presented as the fatty acid fractions of the sum of one per lipid class). Descriptive analytics was applied to verify the data distribution. The summarized lipidome variables were assessed with respect to collinearity, and collinear variables (Pearson's $r^2 > 0.8$) were excluded.

(c) Microbiome diversity was quantified by the Shannon index (Shannon, 1948), which accounts for the relative abundance of each bacterial taxon in the microbiome profile of a participant (Morgan and Huttenhower, 2012), using the `R` package `vegan` (v2.5-6) (Dixon, 2003).

5. **Statistical methods for problem solving:** The abstracted epidemiological, lipid and microbiome profiles were integrated using `R` and analyzed using appropriate statistical models.

   A linear model was used for variable selection, effect estimation and testing for possible interactions with the intestinal microbiota diversity. The model was applied in three variants: crude, sex- and age-adjusted, and lifestyle-adjusted. The model was used for the cross-sectional association analyses (i.e., T3) of weight status and CRP levels with plasma lipidome including an interaction term between each plasma lipidome and the Shannon index. For the longitudinal association analyses, the differences between T3 and T1 (i.e., in regard to lipidome and microbiome variables, and the outcome variables) were used in the linear model, while for the epidemiological covariates, T1 values were used. The model was believed to account for both T3 and T1, and therefore it was not adjusted for lifestyle (i.e., on the basis of the aforementioned data-driven clustering).

   A step forward selection method was applied to select the final model, starting from a model adjusting for age and sex, and ending with the all predictors using Shannon index as interaction term. The stepwise regression method (Venables and Ripley, 2002; Bruce and Bruce, 2017), which is implemented in `bootStepAIC` (v1.2-0), assesses the Bayesian information criterion for model selection. Model variability was investigated and 95% bootstrap confidence intervals were calculated (`R = 10`). Model p-values were adjusted according to Bonferroni correction (Holm, 1979). As well, multiple comparisons using single-step procedure for simultaneous tests for general linear hypotheses were performed on each model as implemented in `multcomp` (v1.4-10) (Hothorn et al., 2008). We considered a significance level of $\alpha = 0.05$.

5.2.3 Evaluation

In this study, the principles for meaningful use of data were applied, and a number of dimensionality reduction approaches were employed to achieve meaningful

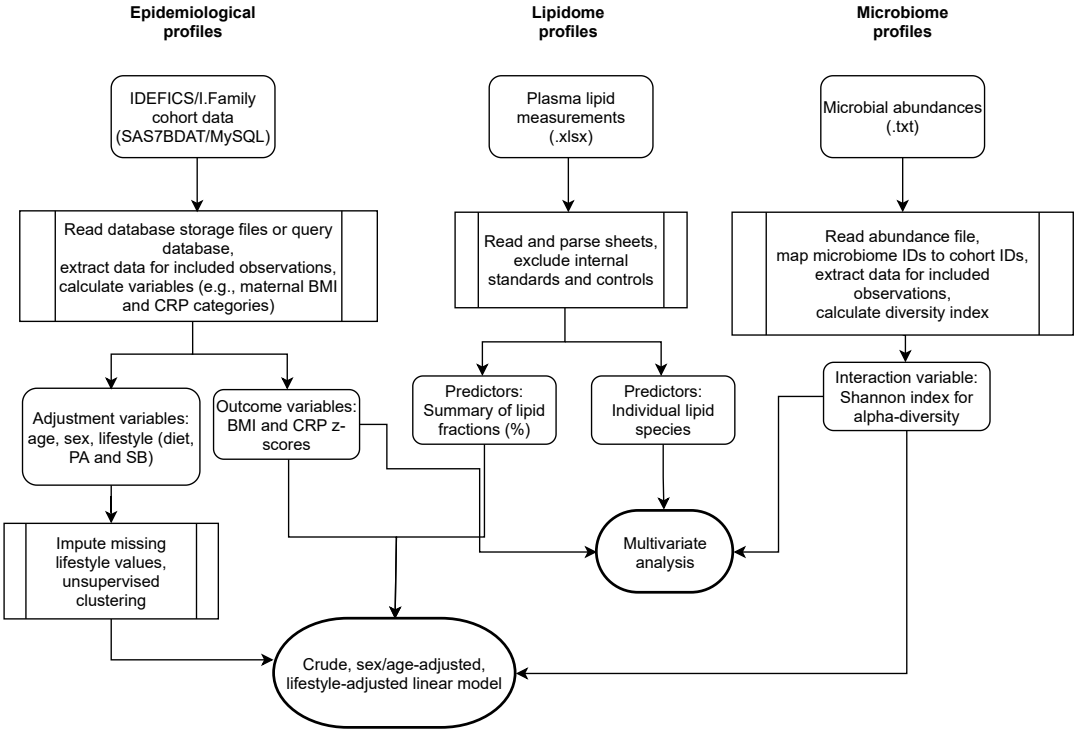


FIGURE 5.2: Data and processes flow diagram for the integration and analysis of epidemiological, lipidomics and microbiomics data. Squares with rounded corners represent the data, processes are shown as rectangular boxes, and terminator ovals represent the statistical analysis. The three profile types are shown in the diagram, including data sources and formats. Data from sources are used as input for extraction and transformation processes. The processes output data either directly to terminators or require further transformations (e.g., handling missing values and clustering). The workflow is simple and flexible; transformed data can be loaded from process to terminator, or temporarily stored on disk if the data security issues allow for that. BMI = Body mass index, CRP = C-reactive protein, PA = Physical activity, SB = Sedentary behavior.

data integration, and, consequently, optimize knowledge extraction. Moreover, solutions for heterogeneous data processing, transformation and integration were implemented. The data integration model proposed here was cumulatively evaluated by the means of statistical methods performance, that is testing for associations between predictors from multiple sources and the outcome. This subsection focuses on evaluating the strategies adopted in data processing and analytics.

Model applicability

Achieving meaningful use of data, and particularly big data integration was presented by Bizer *et al.* (Bizer et al., 2012) as a multi-disciplinary challenge. The integration model elegantly showed the position of data preparation and analysis in the data and process flow. It also showed the importance of problem definition and entity resolution. Entity resolution seems an important step when retrieving a large data set particularly from a database management system. This case study shows

that it is an integral part of meaningful data integration regardless of data set size or structure. Moreover, the generic nature of the model made it attractive for use in a future similar pilot study aiming at integrating fecal lipidome data of the participating children. This is particularly important as fecal samples collection is a non-invasive procedure, which is important in children cohort studies. Moreover, fecal lipidome reflects intestinal function and disease, which, in turn, is known to be linked to various immunological morbidities (e.g., allergy) (Gregory et al., 2013).

The role of statistics

The case study shows that the role of statistical methods implies analytics and effect estimation, data verification and abstraction, and evaluation of data usefulness. Consequently, statistical methods help assess the data integration model effectiveness. First, utilizing single imputation, descriptive analysis and Pearson test for correlation aided the incorporation of informative variables. Second, utilizing dimensionality reduction approaches (i.e., construction of single variables for lifestyle and microbiome diversity) helped minimizing the risk of overfitting by the statistical model. Third, statistical model performance is proposed to be used to evaluate integration effectiveness (discussed in 5.2.4).

Optimizing data acquisition and processing

Data processing was challenged by data heterogeneity, and principles for heterogeneous data processing were applied accordingly (Wang, 2017). Depending on the types of heterogeneity expected, appropriate data processing principles need to be incorporated for optimal data integration. To further optimize data acquisition, data were retrieved from source to minimize self- or past-dependencies. Self- and past-dependency issues in data acquisition practices are not uncommon, specially when dealing with shared data repositories. These issues may lead to inconsistencies in data acquisition and analytics upon, often automatic and regular, updates of the data source. Moreover, these issues promote the need for data transfer, a practice that often violates data flow security. In this study, a single `R` package was developed to encapsulate all curation and analytics employing past/self-dependencies minimization concepts and promoting reproducibility.

5.2.4 Critical appraisal

Evaluation of meaningful data integration

As discussed in this chapter, to achieve the aims of modern cohort studies, purposeful integration of heterogeneous data types is required. This case study argues that dimensionality reduction is a key step in such integration. Bizer *et al.* (Bizer et al., 2012) focused on big data integration. In one of the presented cases, completeness and consistency of the curated data were considered as success metrics.

Nevertheless, challenges of multi-omics data integration were not discussed. A plethora of dimensionality reduction approaches have emerged and were compared (van der Maaten et al., 2007). Those approaches were employed for integrative analysis of multi-omics data (Meng et al., 2016), and their performance was evaluated on simulated and real data (Fanaee-T and Thoresen, 2019). Evaluation approaches and success metrics are not well-developed in this case study.

The application of statistical approaches for data verification and abstraction motivated the proposed evaluation criteria to better quantify integration efficiency. For instance, a systematic adoption of performance measures to evaluate the data integration model is proposed. In particular, at the data acquisition-side, data comprehensiveness can be evaluated looking at descriptive statistics. At the analytics-side, to evaluate the effect of data-driven clustering as an adjustment variable on effect estimation of the statistical model, cross-validation can be utilized.

Multivariate analysis of (multi-)omics data

In this case study, a comparison between the analysis of the abstracted profiles by linear regression and the analysis of the complete profiles using multivariate analysis is required to better evaluate the linear model in regard to variable selection. Multivariate analysis methods are considered the most common for integration and statistical analysis of (multi-)omics data (Huang et al., 2017). Those methods are reviewed (Orešič, 2009; Worley and Powers, 2012; Paliy and Shankar, 2016) and comprehensively compared (Acharjee, 2012), in particular for small sample sizes (Kirpich et al., 2018). The criteria for selecting a multivariate analysis method include: 1) the research question (e.g., variable selection, prediction and classification or discrimination), 2) the number of data blocks (single- vs. multi-omics data), 3) the need for adjusting for confounding variables, and 4) the number and type of response variables (i.e., single or multiple, nominal or continuous). One of the most popular analytical methods in this context is partial least squares (PLS) (Wold et al., 2001), which is used in microbial abundance analysis (Paliy and Shankar, 2016) and lipidomics (Checa et al., 2015; Mundra et al., 2016). In addition, novel methods based on machine learning [e.g., block forests (Hornung and Wright, 2019)] or penalized regression [e.g., priority lasso (Klau et al., 2018)] are being developed and used for predicting clinical outcomes from high-dimensional multi-omics data. Moreover, the interplay between the different associated lipid classes and the related genes and proteins could be investigated through mapping the variables to human metabolites databases (e.g., KEGG) and lipids database (LIPID MAPS) (Cotter et al., 2006).

5.3 Concluding remarks

Modern design of cohort studies aim at advancing our understanding of diseases, and driving therapeutic decision making and drug development. Currently, large data volumes are generated from extensive multi-omics phenotyping of the individuals. Integration of heterogeneous omics data becomes indispensable for study population profiling. Such integration, however, augments the resources needed for sample collection, data acquisition and management, and data analysis.

This case study aimed at testing the applicability of dimensionality reduction approaches in achieving meaningful integration of heterogeneous high-dimensional biological data. In this chapter, I followed a model for meaningful data integration that illustrates the effect of data transformation, abstraction and analytics in improving the usability of multi-omics data and optimizing knowledge extraction. First, the model is flexible; it supports heterogeneous data integration, and accounts for the various aspects of data processing. Second, the application of appropriate statistical approaches improves data integration and usability.

As seen in Chapter 4, there is a need for scalable analytics solutions to accommodate high-dimensional data. Chapter 5 argues that the rapid rise of multi-omics data availability makes such need imminent. Even though current methods for multi-omics data analytics seem suitable for analyzing a large number of predictors, as the number of observations increase, computational and perhaps methodological challenges arise (e.g., due to missing values and imbalanced data). Moreover, integration of heterogeneous data and harmonization of data from different sources can present a bottleneck towards optimal knowledge extraction. The more unstructured the data are, the more challenging harmonization and therefore usability become. Digitalization of the health care system and the rise of the Internet of Medical Things (Dimitrov, 2016) would further drive the development of data integration. The future of primary biological data might benefit from utilizing emerging approaches that support integration, such as data virtualization (Pullokkaran, 2013; Wang, 2017), which provides solutions for data transformation and analysis of heterogeneous data in place and in real-time.

The increasing speed of primary data availability no longer seems a distant future. At the end of the year 2019, the world has embarked a global crisis; a threat that is of pure biological nature. The respiratory disease COVID-19, caused by the novel coronavirus SARS-CoV-2, has triggered a sequence of political, computational and social challenges worldwide. Various resources have been made available for the scientific community to tackle COVID-19 including virus[2] and host (i.e., the UK Biobank (Sudlow et al., 2015)) genetic data as well as computational resources[3]. Open science and sharing of data and resources provide hope to join forces in the face of the pandemic that claimed the lives of 729,393 persons worldwide [Johns Hopkins Coronavirus Resource Center; August, 10, 2020][4]. Similarly,

crowd-sourcing in 2011 fast-tracked the genome analysis and decoding of the *E. coli* O104:H4 (STEC O104:H4) strain that was responsible for the outbreak in Europe (Rohde et al., 2011).

In epidemiology, each data point matters, yet the data points are as good as their usability (i.e., including aspects of secure access, cleaning, integration, and analysis). The application of data engineering principles can drive knowledge extraction through improving data usability. In the final chapter, I summarize the lessons learned from the presented case studies, and highlight the potential impact and challenges of resources and data sharing.

[2]CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. Singer BJ, et al. Preprints 2020, 2020060225. URL: http://cov-glue.cvr.gla.ac.uk/ and The COVID-19 Data Portal. URL: https://www.covid19dataportal.org/

[3]Open-Access Data and Computational Resources to Address COVID-19, NIH. URL: https://datascience.nih.gov/covid-19-open-access-resources

[4]Johns Hopkins Coronavirus Resource Center. URL: https://coronavirus.jhu.edu/

Chapter 6

Conclusions and Outlook

It is profound that numbers (i.e., data) have no way of speaking for themselves; we often speak for them[1]. Data engineering provides not only the tools for making the most sense of the data, but also the recipe for knowledge extraction and for translating the data into actionable predictions.

This thesis addresses challenges in biological data preparation and transformation for analytics, particularly variable selection. The thesis presents custom-made solutions for four areas of application on primary and secondary biological data in health and environmental research. The presented case studies highlight similarities and dissimilarities in biological data handling with respect to the data source (primary *vs.* secondary), type (sequence-based *vs.* relational) and field of application (health *vs.* environment). In this final chapter, I discuss the outcome of the four case studies and its relevance to the needs of biological data curation and analytics. I also give an outlook into the needs of the future landscape of biological data.

6.1 The making of knowledge in health and environment

Knowledge extraction from primary high-dimensional sequence data

Processing and analyzing observation sequence data are routine activity in biological research, with the purpose of knowledge extraction. In Chapter 2, I focused on transcriptomic studies in environmental research, a major valuable source of biological primary sequence data. Transcriptomic studies provide a glimpse on the metabolic potential of complex organisms growing under harsh conditions, bypassing the obstacles of genome sequencing of environmental samples. However, transcriptomic data are high-dimensional and generated by high-throughput sequencing platforms. Existing workflows are often successfully used for knowledge extraction, mainly through the utilization of a customized multi-step dimensionality reduction protocol, which I explored using the case study in Chapter 2. I utilized an existing workflow (i.e., Trinity full-suite solution), which seamlessly handled data processing and quickly provided easy-to-interpret results. Nevertheless, the large

[1] Silver N. The Signal and the Noise. New York, US: Penguin Books; 2015. Reprint edition.

number of transcripts produced by Trinity required a mixture of statistical methods and programming tools for dimensionality reduction. I employed those tools and methods even prior to transcript identification, through the steps of quality control by filtering low-quality short reads and unannotated transcripts, and the differential expression analysis by using k-mean clustering to infer patterns of gene expression across time and scripting-based consolidation of functional annotation data.

The case study highlighted an issue when analyzing high-dimensional primary sequence data from a non-model organism, namely effective gene characterization. Gene characterization is necessary to pinpoint the genes responsible for the species' metabolic activity under changing conditions. However, the under-representation of the studied group in public databases and the lack of information on their genomes and splicing events affect accurate gene characterization. Therefore, knowledge extraction from primary high-throughput sequence data is largely based on domain knowledge and customized dimensionality reduction approaches, in particular using ontologies at gene-level instead of transcript-level, k-mean clustering of differentially expressed genes, and pathway analysis.

Scalability in handling secondary sequence data

Large volumes of high-quality observation data collected on individual or communities of organisms through global projects are giving rise to valuable secondary sequence data and encouraging information integration and meta-studies. Scalability, modularity and reproducibility are essential for such data-driven studies, offering flexibility for information integration, and accounting for the diversity of organisms to be included, the differences in sequencing technologies, and possible integration schemes. In Chapter 3, I explored the potential for improving the scalability of the Trinity workflow to achieve information integration and analysis of secondary transcriptomic environmental data. Therefore, I designed and implemented a pipeline to acquire, integrate, and analyze gene expression data archived in public repositories. To test the pipeline usability, I also designed a meta-analysis case study. To achieve satisfactory scalability, modularity and reproducibility for integration and analysis of the gene expression data, I focused on automation at two levels: dynamic acquisition of annotation databases to update gene characterization results as needed, and automation of the workflow steps.

The case study highlighted several bottlenecks in secondary sequence data analysis, such as hardware requirements and the potential for flexible computing plans (e.g., cloud computing), experimentation with more sophisticated analyses (e.g., based on information theory) which requires scalable data transformation, and data ownership. Interestingly, the technical bottlenecks also challenge handling non-sequence-based secondary data as addressed in Chapter 4, summarized below.

Integration and statistical modeling of secondary high-dimensional relational data

Bottlenecks in integration and statistical modeling of secondary high-dimensional relational data directly impact modern epidemiological research. The steep rise in routinely collected health data (i.e., electronic health care databases) and the use of common data models both promote the use of such valuable data in pharmacoepidemiological research. For instance, the data could be used for monitoring drug safety in large populations in the post-marketing phase. When paired with the utilization of molecular-based ontologies, analysis of secondary structured relational epidemiological data could better explain the underlying mechanisms of disease outcomes in the light of the ever-growing body of molecular biology knowledge. However, similar to secondary sequencing data, limitations in performance and scalability of analytics arise. In Chapter 4, I used a large-scale signal detection case study to address aspects of both data integration and scalability of current implementations of specialized statistical methods. Therefore, I optimized and developed scalable portable solutions for acquisition and transformation of molecular knowledge data from online data sources (`SQLite` database), transformation of high-dimensional relational health care claims data in conjugation with the molecular knowledge (utilizing on-disk intermediate storage objects to better scale to big data volumes), and adapting statistical methods for high-dimensional data and high-performance computing (HPC) resource (utilizing parallelization and minimization of memory fragmentation).

The case study showed how both the data source (ontology data) and dimensions (longitudinal data) influenced acquisition and transformation strategies. The case study also showed how the transformed data dimensions drove both analytics choice (e.g., due to limitations of group-based penalized regression implementations) and implementation (i.e., the adaptive rank truncated product for outcome prediction). Finally, the case study highlighted the need for exploring benchmarking utilizing simulated data to systematically investigate the effect of group/block number and/or size on analytics, tools augmenting relational database management systems for large-scale data processing and analytics, and staging, to design and utilize a relational middle layer between health care claims databases and analytics (e.g., a column-oriented database management system) to better support transformation of high-dimensional data.

Integration of heterogeneous primary epidemiological data

It may seem that a model for integration of data from multiple sources is strictly required for large-scale studies. However, successful data integration, regardless of the data volume, requires a model for meaningful use of data to distinguish between informative and irrelevant attributes, and to achieve optimal knowledge extraction. As primary data remain the largest contributor to the biological data sphere, in

Chapter 5, I explored the principles of meaningful use of data, and applied them in a case study to integrate heterogeneous data from a cohort pilot study (i.e., epidemiological, lipidomics and microbiomics data). I adopted a flexible model that supports meaningful integration of heterogeneous data, starting from clear problem definition, database querying, transformation of heterogeneous data, statistics-based entity resolution using dimensionality reduction approaches (e.g., correlation-based verification and data-driven clustering), and problem solving (i.e., optimized linear model-based variable selection).

The case study showed the need for adopting a flexible data model for handling primary heterogeneous data regardless of the sample size. It also showed the importance of applying heterogeneous data processing practices on modern cohort study data. The case study highlighted the role of statistical methods in both entity resolution and problem solving, as the application of appropriate statistical approaches improves data integration and usability. Evaluation of meaningful data integration metrics are yet to be adopted for multi-omics data. Moreover, evaluation of extracted knowledge is required, for instance, comparing two analysis approaches: linear regression of the abstracted profiles and multivariate analysis of the complete profiles.

6.2 In a data-driven new world

Biology is a unique core contributor to world data, either through research (epidemiological studies and ecological batch experiments), environmental surveillance or health care. This is life's data, in health and disease, in disaster and prosperity, in prediction and prevention. Biological data are expensive, yet the real cost of biological data is mush higher than the sum of acquisition and management costs. The real cost includes that of data storage, security, (pre-)processing, linkage, and analytics. The pathway from data to information is long, even for standardized data (e.g., medical imaging data), which explains the high cost of data transformation into analysis-ready data. Nevertheless, the cost falls short in representing the value of biological data. Every point's meta-, raw-, and intermediate data is invaluable as it drives inference and prediction forward, and requires data protection, which adds a further layer of complexity to biological data handling. Thus, the value and the growing volume of the data characterize modern environmental and epidemiological research data, both are characteristics of big data. In this section, I discuss the data types that this thesis addressed in the light of big data characteristics. I also highlight a number of overarching needs of the biological data processes and possible solutions.

6.2.1 Big data and its characteristics

Mauro *et al.* formally defined big data as the representation of "the information assets characterized by such a high **volume**, **velocity** and **variety** to require specific

technology and analytical methods for its transformation into **value**" (De Mauro et al., 2015). This definition captures the five pillars of big data: volume (i.e., storage requirements), velocity (speed of generation and processing), variety (of data types which requires data fusion), veracity (as in data quality and reliability), and value (of extracted information). The definition also implies the need for data transformation, processing, and advanced analytics to ultimately achieve information generation and decision making. Such lengthy costly pathway to knowledge characterizes big data. A generic characterization of big data, therefore, became: the data that cannot be handled within the resource constraints on a single machine, where constraints for transformation and information extraction are time, memory or disk space.

It is not possible to discuss big data without considering a major contributor to it, which is unstructured data. Unstructured data, human-generated (e.g., textual) and sensor-generated (e.g., imaging) data, conform to no known data model and cannot be stored in or processed by a database management system (Buneman et al., 1995, 1996). The value of data depends on the knowledge it generates and actionable outcome derived from it (e.g., environmental management). The value of big unstructured data collected for general observatory purposes (e.g., ecological surveys) was argued to be less than that of standardized monitoring data, which are collected for a particular purpose in generating knowledge (Bayraktarov et al., 2019). Bayraktarov *et al.*, therefore, argue for the utilization of benchmarking against high-quality data, identification and maintenance of key time-series datasets, and investment in data curation and sharing as in data collection (Bayraktarov et al., 2019). On the analytics-side, mining unstructured (bio)medical data could contribute to the identification of drug-disease relationships in secondary (i.e., bibliographic) repositories (Ji et al., 2015). This thesis did not deal with unstructured data. Nevertheless, the non-sequence epidemiological data from the case studies in Chapters 4 (e.g., dispensations data) and 5 (e.g., wearables data) were originally unstructured; the data were validated and transformed into structured relational data through highly standardized procedures.

Addressing cases of real-world data, the thesis inspires three questions. First, when to consider biological data big data? Second, would such a consideration influence the practices needed for transformation into analysis-ready data and the analytical methods used? Third, is the relationship between the data dimensions and the knowledge gained a linear association; could using bigger data lead to gaining more knowledge? I reflect on these questions below.

6.2.2 In big and small: The path to knowledge

Having a large volume (e.g., sequence data), being produced at a high velocity (e.g., high-throughput omics data and routine health care data), variety (requiring integration), veracity, and being of utmost value, the majority of biological data would

be considered big data. Such consideration would motivate the transfer of big data handling practices from other fields (e.g., image analysis and data mining).

Concerning the case studies presented in this thesis, regardless of the data dimensions (large-scale *vs.* pilot) and type (sequence-based *vs.* relational), transformation into analysis-ready data as well as variable selection-relevant analytics were complex. Even in "small" studies (Chapter 2 and 5), the path to analysis-ready data and knowledge required the utilization of a data model for data transformation (and integration), and high computational requirements for analytics due to the high-dimensional nature of omics data. Both requirements were clearer in the large-scale studies (Chapter 3 and 4).

An obvious aspect of big data is the high computational requirements for data transformation and analysis. Genomic data, in particular, is accounted for as big data, as one whole human genome produces over a 200 GB of raw and analysis-ready data generated by HPC facilities. Genomics is a main driver of the recent advances in personalized medicine, which is translated into drug development. In particular, 42% of the new drug approvals by the U.S. Food and Drug Administration (FDA) in 2018 were personalized medicines[2]. Volume, velocity (rise in sequenced genomes) and value of genomic data motivated the utilization of big-data-relevant technologies such as cloud computing (e.g., AWS Genomics) and GPU-accelerated computational framework for genomics (e.g., by NVIDIA).

In addition to genomic data, although highly structured, electronic health care data are currently considered big data (Andrews et al., 2014; Umemoto et al., 2019), with respect to three factors: 1) the rapid growth in data volume and value, 2) data fusion from different sources, and 3) data analysis in large-scale studies. Using multiple data sources would require: 1) development of a unified data model, 2) utilization of ETL workflow or alternatives to populate the model with the data, and 3) deployment of analytical workflows on HPC resources for such large-scale studies (Curcin et al., 2008). From the analytics perspective, the standard methods for risk prediction fall short in utilizing the full spectrum of big data. Therefore, the alternatives include applying data mining and machine learning methods, and adapting epidemiological methods (e.g., penalized regression) for HPC platforms. Therefore, it seems that computational requirements are the tip of the iceberg, further needs are highlighted below.

1. **Data model for acquisition, curation and integration of heterogeneous high-dimensional data** Although implicit, the first step in handling either sequence (Chapter 3) or relational data (Chapter 4) is the creation of a data model for transformation and integration, which is an integral step to support application of target analytics. Particularly when handling heterogeneous data (as

[2]Personalized medicine at FDA: A progress & outlook report, Personalized Medicine Coalition. URL: https://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/PM_at_FDA_A_Progress_and_Outlook_Report.pdf

in Chapter 5), careful consideration and creation of a suitable data model are required to represent the relationships between the data elements as they are in real life. Currently, data models and schemas to support studies using secondary data from multiple electronic health care databases (Pacurariu et al., 2018) and meta-omics sample data (Rambold et al., 2019) have been developed. The use of those models could be propagated for data handling and transformation.

2. **Empowering predictive analytics filling staging and production gaps** The case study in Chapter 4 pinpoints areas for empowering analytics. First, a health care claims databases could benefit from (or even offer as a service) an attached staging area, a middle layer between the database and analytics, such as a column-oriented database management system (e.g., MetaKit). Establishing a two-warehouse data management system, one for original data and another for the transformed analysis-ready data would offer a solution for large-scale studies challenges. Second, the difference in software development practices between R and Python, where Python is more production-oriented is apparent. Python (and Julia) are also better suited for high-dimensional data. However, regression-based methods (e.g., Group LASSO and conditional LASSO), successfully implemented in R, are yet to be implemented in Python or other platforms including discipline-independent database-based solutions for scientific large-scale data analytics (e.g., SciDB). This is an opportunity for implementation of analytical methods suited for high-dimensional health data.

3. **Evaluation of extracted knowledge** The ability to evaluate the data transformation approaches and their effect on the information lost (or gained) is crucial. Benchmarking against state-of-the-art methods (or of novel methods using reference sets), and simulations are possible directions. However, it is certainly not feasible to compare methods/datasets in every study to evaluate extracted knowledge.

4. **Data sharing: The COVID-19 test** Data sharing is an integral topic when discussing research data, personalized medicine and ecology, particularly through the FAIR principles (Wilkinson et al., 2016). In 2020, data sharing had become "vital" due to the coronavirus disease 2019 (COVID-19) pandemic. Not only large numbers of virus and host sequences are publicly available, but also patient-level epidemiological data (e.g., electronic health records, physiology, laboratory, imaging, and treatment data) are recorded, yet these patient data are not suited for sharing (Cosgriff et al., 2020). It is, therefore, argued for the need for a multinational COVID-19 electronic health record database (Cosgriff et al., 2020) to facilitate the application of sophisticated analytics (Cosgriff et al., 2020; Peiffer-Smadja et al., 2020). As well, a number of solutions have been suggested to address data heterogeneity and security issues (Paul and Chatterjee, 2020). In addition to (bio)medical data, data from

wearables (e.g., smartwatches) have been collected, compared and evaluated for pre-symptomatic detection of COVID-19 (Mishra et al., 2020). The COVID-19 pandemic has been testing the the current capacities for integration, sharing and analytics, with billions of lives are at stake. These efforts will certainly benefit respiratory and heart disease monitoring, personalized medicine (Denny and Collins, 2021). The deep understanding of the importance of data sharing and availability motivated the national initiative in Germany, the National Research Data Infrastructure (NFDI), aiming at managing scientific and research data, in terms of storage and accessibility at the national and international levels. Two consortia are relevant to the case studies in this thesis, NFDI4Health[3] and NFDI4Biodiversity[4]. NFDI4Health aims at providing a central registry for health data and metadata, analytics software, and data linkage services. NFDI4Biodiversity focuses on serving ecological research, facilitating access to modern technologies and a comprehensive repository of environmental data.

6.3 Outlook: When life depends on it

This thesis is a glimpse into the data of life, from health and environment. The data are of extreme value, if integrated and interrogated properly. Ultimately, the data drive preparation and transformation, which consequently directly influence data analytics choice and performance, and influence extracted information, calling for customized solutions for every case. Nevertheless, a number of challenges remain such as cost (e.g., storage and computation), security and data sharing. As life itself depends on it, acquiring high-quality data must be completed with the creation of a suitable data model and necessary computation requirements for transformation analytics. Investing in scalable solutions is inevitable. As a "small" database grows in volume and value, a shift in perception must be considered to address hard- and software requirements for transformation, warehousing and analytics of a soon-to-be an "extensive" repository. Integration as well is foreseen in the future. Perhaps soon, health records, environmental microbial data and particle pollution data would be integrated for air pollution monitoring and decision making, reminding us of the interconnected world that we can take action to preserve.

[3]NFDI4Health - National Research Data Infrastructure for Personal Health Data. URL: https://www.nfdi4health.de/en/

[4]NFDI4Biodiversity. URL: https://www.nfdi4biodiversity.org/de/

Appendix A

Technical Supporting Material

A.1 Sequence similarity and E-value

Sequence similarity searches have been used in biology for decades, and have many applications. These applications include: 1) annotation (i.e., finding homologs of sequences of interest (nucleotide or amino acid sequence) in public data repositories), 2) inferring the evolutionary origin as to: a) identify homologs of sequences that share statistically significant similarity with (i.e., identify orthologs that descend from a common ancestor), b) identify homologs of sequences that share statistically significant similarity within the same organism (i.e., identify paralogs, gene duplication events).

BLAST, the most common and well-established sequence comparison algorithm, uses four steps for sequence comparison (Altschul et al., 1990; Kerfeld and Scott, 2011). These steps [according to (Kerfeld and Scott, 2011)] are: 1) It chops the query sequence provided by the user into "words", and accounts for mutations in these words by creating a list of synonyms for each word. These words and synonyms are then scored according to their similarity to the query sequence based on BLAST's substitution matrices. 2) BLAST scans the entire database for sequences that contain these words and their synonyms. 3) It then moves forward creating an alignment between the query and the "subject sequence"; if the score (S) of this "ungapped" alignment is high, the query and the subject sequences are considered homologs/similar. Gaps in the alignment represent the insertion/deletion of an amino acid or a nucleotide. The score drops (is penalized) based on gap existence/opening and gap extension. 4) The alignment is terminated once the match score drops below a predefined threshold score, therefore, the alignments are called "local" as compared to "global".

The aforementioned raw scores S are later normalized across different penalties and matrices used into S' (bit score). Each subject will have an S' value to reflect its similarity to the query. S' is further normalized to account for the size of the database against which the search took place (n in residues; amino acids or nucleotides) and the length of the query m; this yields E, with $E = {}^{(n \times m)}/_{(2^{S'})}$. E represents the

number of subject sequences that a BLAST search is expected to retrieve by change alone from a database (that size), where those subject sequences have an S' larger than or equal the S' calculated from the match/alignment. In case the query and the subject sequence "hit" are very similar, E would be small and would reflect the confidence that these two sequences are homologs. E for a BLAST search using the same query can change over time due to the change in database size.

A.2 Integration and modeling of secondary health data: Setup, methods and results of a simulation study

Contribution: R. Foraita, L. Dijkstra and I planned the concept. I conducted the literature research, data curation, and pipeline setup with R. Foraita. Statistical methods were compiled by R. Foraita and myself, and revised by L. Dijkstra. The appendix is part of a project deliverable report that I, L. Dijkstra and R. Foraita wrote.

Preface

In order to reduce the number of patients affected by adverse drug events (ADE), it is of utmost importance to identify patient groups that are at risk. Drug exposures and comorbidities often form the basis for creating patient risk profiles. Instead of basing the profile on the associations between individual drugs and diseases with the ADE, we propose to use domain knowledge by linking drugs, diseases and the ADE to so-called functional targets (FTs), i.e., a pathway of interacting biomolecules (e.g., receptors). Here, we compare ten statistical methods that are able to exploit this underlying group structure to 1) infer the FTs that affect the ADE risk, and 2) predict whether a patient will suffer the ADE given his/her drug exposures and diseases. The FTs are curated from the online database Kyoto Encyclopedia of Genes and Genomes (KEGG). The methods' performance is assessed based on simulated health care claims data. The area under the receiver operating characteristics curve (AUC) is used as performance measure. The adaptive rank truncated product (ARTP), a gene set enrichment strategy from the field of genetic epidemiology, performed best in most parameter settings for inference and prediction.

Annotation of drugs and diseases based on FTs using KEGG

There are several public repositories of largely manually curated biological and chemical databases that link drugs and diseases to FTs. Examples include the Therapeutic Targets Database (TTD) (Yang et al., 2016), ChEMBL (Gaulton et al., 2012), Search Tool for Interacting Chemicals (STITCH) (Szklarczyk et al., 2016), and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2017). For this simulation, we consider KEGG as a comprehensive publicly available online database that allows linking drugs and diseases to the same FT, comprising various biological processes, components or structures with which drugs and diseases interact. Most importantly, the data in KEGG are cross-referenced with the Anatomical Therapeutic Chemical (ATC) classification system and International Classification of Diseases (ICD), which makes it easier to associate the FTs with the drugs and diseases stored as real health care claims data. For these reasons, we use the information in KEGG to map human diseases and approved drugs to FTs.

Methods

We compare several statistical methods with respect to their ability to: 1) infer the groups that effect the ADE risk, and 2) predict whether a patient will experience the ADE based on his/her drug exposures and comorbidities. Table A.2.1 gives an overview of all methods. Note that some of the methods are only suitable for prediction, see column 'I' (inference) and 'P' (prediction) in Table A.2.1.

Notation

First, we introduce some notation that will be used throughout this appendix. Let us suppose that we have m covariates $x_i = (x_{i,1}, \ldots, x_{i,m})^\top$ with binary entries for each patient $i = 1, 2, \ldots, n$, comprising all drugs and diseases in the data set. In addition, we have a binary response vector $y \in \{0, 1\} \in \mathbb{R}^n$ that denotes the occurrence of the ADE.

The relationship between $\mathbf{X}$ and y is modeled by the following logistic regression model:

$$\log \left\{ \frac{P(Y = 1 \mid \mathbf{X} = \mathbf{x})}{1 - P(Y = 1 \mid \mathbf{X} = \mathbf{x})} \right\} = \eta(\mathbf{x}). \tag{A.1}$$

The function η is the linear predictor

$$\eta(\mathbf{x}) = \beta_0 + \mathbf{X}\beta^\top$$

where $\mathbf{X} = (x_{i,j})_{i,j}$ denotes an $n \times m$-dimensional matrix of covariates with $x_{i,j}$ being the j-th covariate of individual i, $\beta_0 \in \mathbb{R}$ is the intercept and $\beta \in \mathbb{R}^m$ is a vector of regression coefficients. Furthermore, each of the m covariates can be assigned to G groups. Each group $g = 1, \ldots, G$ has a group size, i.e., s_g. Groups represent FTs and can, thus, possibly overlap, which means that some covariates can be assigned to multiple groups simultaneously. Let the matrix $\mathbf{X}^{(g)}$ represent a submatrix of the design matrix $\mathbf{X}$, where the columns correspond to the covariates contained in group g. The coefficient vector for this group is denoted by $\beta^{(g)}$. The corresponding linear predictor of group g can then be written as:

$$\eta\left(\mathbf{x}^{(g)}\right) = \beta_0 + \sum_{g=1}^{G} \mathbf{x}^{(g)\top} \beta^{(g)}. \tag{A.2}$$

Regularized regression models

Regularized regression methods exploit sparsity to detect signals in particularly high-dimensional data sets. Hence, they might be an attractive approach for signal detection when only a minor proportion of all drugs on the market could cause the ADE of interest. Regularized regression methods estimate a vector of regression

coefficients β by minimizing an objective function $S(\beta)$ composed of a loss function L that assesses the deviance between the outcome and the linear predictor in combination with a penalty $P(\beta \mid \lambda)$:

$$S(\beta) = L(\beta_0, \beta \mid y, \mathbf{X}) + P(\beta \mid \lambda),$$

for some $\lambda \geq 0$. The penalty $P(\beta \mid \lambda)$ regularizes the parameter estimation by tuning the parameter λ to control both coefficient shrinkage and variable selection. The most popular regularization method, the least absolute shrinkage and selection operator [*lasso*; (Tibshirani, 1996)], minimizes the negative log-likelihood along with the l_1-penalty to shrink the coefficients towards zero with some coefficients set to exactly 0:

$$\underset{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^m}{\text{minimize}} \Big\{ -\frac{1}{N} L(\beta_0, \beta \mid y, \mathbf{X}) + \lambda \|\beta\|_1 \Big\}$$

where $\|a\|_1 = \sum_i |a_i|$ and the log-likehood function of a logistic model as formulated in equation (A.1) takes the form

$$L(\beta_0, \beta \mid y, \mathbf{X}) = \sum_{i=1}^{n} y_i \eta(\mathbf{x}) - \log\left[1 + \exp\{\eta(\mathbf{x})\}\right].$$

One drawback of the lasso is its overestimating behavior. The *adaptive lasso* (Zou, 2006) addresses this issue by decreasing the bias and, hence, reducing the number of false positives. It is a two-stage procedure, where, in the first stage, pilot estimates $\tilde{\beta}$ are obtained, which are then used to re-weigh the regression coefficients of a lasso regression in the second stage. In this simulation study, we apply *ridge regression* (Hoerl and Kennard, 1970) in the first stage.

Nevertheless, both methods are not able to include prior knowledge about group structures. In order to include this prior knowledge, we propose a very simple strategy, which we will refer to as the *naïve group lasso* (NGL). First, for each group, we determine a 'group variable' that reflects the covariate values of that particular group. The lasso is then applied to these constructed group variables, rather than to the individual covariates in order to identify groups possibly associated with the ADE of interest. There are two approaches in which we define these group variables. The first approach defines the group variable as the sum of the covariates within that group, i.e., $x_{\text{sum}}^{(g)} = \sum_{j=1}^{s_g} x_j^{(g)}$. We will refer to this definition as the *sum* approach. The second approach, to which we will refer to as the *any* approach, summarizes the group by setting the group variable to 1 when any of its covariates is 1, and 0 otherwise. The latter approach is intended to reflect the hypothesis that drugs from the same FT might lead to the same ADEs due to their similar chemical properties.

In addition, there are various other regression methods that put additional regularizations on group membership by using different penalty functions. Yuan and Lin (Yuan and Lin, 2006) proposed the *group lasso* to select entire groups of covariates which solves the convex optimization problem

$$\underset{\beta_0\in\mathbb{R},\beta^{(g)}\in\mathbb{R}^{s_g}}{\text{minimize}} \left\{ -\frac{1}{N} L\left(\beta_0, \beta^{(g)} \mid y, \mathbf{X}^{(g)}\right) + \lambda \sum_{g=1}^{G} w_g \|\beta^{(g)}\|_2 \right\}, \tag{A.3}$$

where the log-likelihood function uses the linear prediction as formulated in (A.2) and $(w_g)_{g\in G}$ are positive weights that account for different group sizes (Meier et al., 2008). As it is common that a particular covariate x_j is included in more than one group, we also investigate overlapping group strategies based on the work of Jacob *et al.* (Jacob et al., 2009). This might be of great importance in certain cases; consider, for example, a situation where drug X belongs to both the functional targets A and B. The association between X and the ADE, however, is driven solely by target A, and the drug's membership to target B is irrelevant. The minimization problem for the *overlapping group lasso* is formulated as

$$\underset{\beta_0\in\mathbb{R},\beta\in\mathbb{R}^{m}}{\text{minimize}} \left\{ -\frac{1}{N} L(\beta_0, \beta \mid y, \mathbf{X}) + \lambda \sum_{g=1}^{G} w_g \|\gamma^{(g)}\|_2 \right\}, \tag{A.4}$$

where $\gamma^{(g)} = \left(\gamma_1^{(g)}, \ldots, \gamma_m^{(g)}\right)^\top$ are latent coefficient vectors satisfying $\sum_{g=1}^{G} \gamma^{(g)} = \beta$ with $\gamma_j^{(g)} = 0$ if x_j does not belong to group g and $\gamma_j^{(g)} \neq 0$ otherwise. Obozinski *et al.* (Obozinski et al., 2011) showed that solving Equation (A.4) is equivalent to solving the following minimization problem with respect to γ where γ consists of all elements of $\gamma^{(g)}$:

$$\underset{\beta_0\in\mathbb{R},(\gamma^{(g)}\in\mathbb{R}^{s_g})_{g\in G}}{\text{minimize}} \left\{ -\frac{1}{N} L\left(\beta_0, \gamma \mid y, \tilde{\mathbf{X}}\right) + \lambda \sum_{g=1}^{G} w_g \|\gamma^{(g)}\|_2 \right\},$$

where $\tilde{\mathbf{X}}$ is an $n \times \left(\sum_{g=1}^{G} s_g\right)$ new design matrix with duplicated columns from group overlapping covariates. With this transformation, the overlap group lasso is equivalent to the group lasso and can be solved by existing and computationally efficient algorithms (e.g.,). This strategy was also applied for the naïve lasso and the group exponential lasso. The *group exponential lasso* [GEL; (Breheny, 2015)] uses the concept of bi-level variable selection to select important groups as well as the important individual covariates in those groups. This is of interest if not only complete groups should be selected but also single variables. Its penalty contains an additional decay parameter to control the degree to which variables are selected together within groups. The different penalty functions that we use in the simulation study are summarized in Table A.2.1. All regularized logistic regression methods employ 10-fold cross-validation to select the tuning parameter, λ, that minimizes the AUC.

Adaptive rank truncated product

The *adaptive rank truncated product* (ARTP) is a gene set enrichment method that was originally designed for single nucleotide polymorphism (SNP) data (Yu et al., 2009). It is a hypothesis testing approach to select biological pathways that are enriched with genetic variants to be associated with a phenotype. The method preserves the correlation structure between genes by using permutation tests, and it has the potential to detect subtle effects of genetic variants in a pathway that might be missed when assessed individually. The ARTP uses p-values from any statistical association test performed between individual SNPs and the disease outcome. Here, a logistic regression model is applied to analyze the relationship between each covariate and the ADE. The resulting p-values of the respective regression coefficients are then used for the ARTP. Since the ARTP handles each group independently, this method can be also applied to overlapping groups. We modified the ARTP to detect associations between ADEs and functional targets when using binary health care claims real or simulated data. A group is selected when the p-value of the respective permutation test is lower than .1. As the ARTP is not used for risk prediction, we adopted the following strategy to apply ARTP for individual risk predictions. We propose, first, to predict $\hat{y}_i^{(g)}$ for each selected group, and, second, to average the individual group risk predictions to achieve an overall prediction, i.e., the individual risk prediction $\hat{y}_i = 1$ if $n^{-1} \sum_{g=1}^{G} \hat{y}_i^{(g)} > .5$, and 0 otherwise.

Block forests

As a machine learning approach, *block forests* (Hornung and Wright, 2019) are a further development of random forests that is able to combine different types of omics data for outcome prediction. Random forests are known to capture complex dependence structures in data, and block forests additionally allow for including *a priori* known group structures in the analysis to improve the prediction performance. This is facilitated by modifying the split point selection procedure of random forests to the group structure in the data. Overlapping group structures can be analyzed without further modifications. The initial implementation of block forests could only be applied for risk prediction not allowing for variable or block importance estimation, and therefore, it is not used for variable and group selection.

Simulation setup

Each simulated data set consists of $n = 2000$ patients and $m = 1000$ binary covariates, $X_1, \ldots, X_{1000}$, that represent both the drugs and the diseases. The outcome is represented by a binary vector, $y \in \{0, 1\}^{2000}$, which denotes the ADE occurrence in a patient. The covariates are independent, i.e., $X_{i,j} \sim$ Bernoulli (p_j) for $i = 1, \ldots, 2000$ and $j = 1, \ldots, 1000$, where p_j is the marginal probability of taking the drug or having the disease. The marginal probabilities are drawn from a Beta distribution with the shape and rate parameter set to 2 and 15, respectively, to reflect that the majority

of drugs and diseases tend to appear rather infrequently. The simulated covariates are either independent or correlated. In case of the latter, an autoregressive (AR-1) correlation matrix is used with a correlation coefficient of $\rho = .25$.

Ten of the 1000 covariates have a causal effect on the ADE, as in that they are truly associated with the ADE. We refer to these 10 covariates as the *causal covariatves* in the following, although, in this study, we only test for association. In case of the causal covariates, the association has either an odds ratio (OR) of 1.5 (weak), 3 (medium) or 5 (strong effect); the other covariates have an OR of 1. The regression coefficient is set accordingly, i.e., $\beta = \log(\text{OR})$. The intercept, β_0, is determined numerically such that approximately 50% of the patients experienced the ADE as a case-control study design.

All covariates are assigned to groups. The causal covariates are distributed over these groups in two ways: 1) each causal covariate belongs to a different group, and 2) empirically five causal covariates are assigned to one group; the other five are assigned to another group. That means: the proportions of truly associated variables in one group are either 10% or 50%. The non-causal covariates are randomly distributed over the groups. The number of covariates in each group, denoted by $s_1, \ldots, s_G$, are randomly drawn from the sizes of the 303 FT groups present in the online database KEGG (see Figure A.2.1 for their empirical distribution). The group sizes are drawn such that their sum is equal to the number of covariates, i.e., $\sum_{g=1}^{G} s_g = 1000$. We sample from an empirical distribution in order to obtain realistic group sizes.

In the aforementioned setup, the groups do not overlap, i.e., each drug/disease belong to one group only. We refer to this as the setting of 'no overlap'. However, covariates might usually belong to several groups. In order to simulate this setting, we randomly select 100 covariates (both causal and non-causal) and assign them to newly created groups. By doing so, these 100 covariates belong to at most two groups simultaneously. The sizes of these new groups are, as before, sampled from the empirical distributionof KEGG groups. We refer to this second setting as the 'overlap' case. Table A.2.2 shows an overview of the 12 parameter settings used in the simulation study. Overall, we simulate 50 data sets for each setting. The data sets are split into a train and test set with 1,320 and 680 observations, respectively.

Software

The simulation study uses the following packages in R (v3.4.3):

- `simstudy` v(0.1.16) for synthetic datasets generation
- `KEGGREST` (v1.18.1) for extracting drug and disease target information from KEGG

- `glmnet` (v2.0-16) for lasso (L), adaptive lasso (AL), naïve (overlap) group lasso (NGL, NOGL)
- `grpreg, grpregOverlap` (v3.2-0 and v2.2.0, respectively) to calculate (overlap) group lasso (GL, OGL) and group exponential lasso (GEL, OGEL)
- our own package for identification of risk groups in pharmacovigilance using penalized regression and machine learning `RGP`[1] to: 1) acquire and transform KEGG data, 2) create synthetic overlapping and non-overlapping grouping structures based on KEGG data, 3) calculate ARTP for inference and prediction, and 4) select those λ that minimizes the cross-validated AUC for GL, OGL, GEL and OGEL
- `blockForest` (v0.2.1) for block forests (BF)[2]

Results

For each of the 12 different parameter settings (Table A.2.2), we generate 50 data sets, and we apply all methods listed in Table A.2.1 to each of these 600 data sets using AUC as a performance measure. Subsection A.2 shows to what extent the methods are able to infer the groups that have a direct effect on the ADE, while the methods' performance in predicting individual risks are presented in A.2. The figures in this subsection omit the results for OR $= 3$. In addition, we only show the results for the AL with $\kappa = 2$, since it performed either best or comparable to the other values of κ.

Inferring groups

Each method suitable for inference, see Table A.2.1, is applied to the train data of each of the simulated data sets. Figures A.2.2 – A.2.5 show the box plots with the resulting AUCs. The AUCs reflect the extent to which the methods are able to infer the groups (i.e., functional targets) that have a direct effect on the ADE. On the one hand, Figures A.2.2 and A.2.3 show the results when there is *no overlap* between the groups, i.e., each drug/disease belongs to one group only. Figures A.2.4 and A.2.5, on the other hand, show the results where there is *overlap* between the groups. The results with each of the 10 causal covariates being in different groups are shown in the Figures A.2.2 and A.2.4. Figures A.2.3 and A.2.5 show the results if half of the causal covariates are assigned to one group, while the other five are assigned to another. The performance of a random classifier, i.e., an AUC of 0.5, is depicted in each plot with a dashed line.

The figures show, as one might expect, a clear performance increase when the effect size changes from OR $= 1.5$ to OR $= 5$. When the drugs and diseases are correlated (see the lower row of the box plots), the performance drops significantly for

[1] Available under GPL-3 license http://www.github.com/bips-hb/rgp

[2] Using `blockfor` function with the parameters: `block.method = 'BlockForest', splitrule = 'gini', nsets = 100, num_treesoptim = 1000`

OR $= 1.5$. In case of OR $= 5$, this is not so clear. Interestingly, when the causal covariates are equally distributed over two groups (see Figures A.2.3 and A.2.5), we observe the opposite trend: the methods' performance improves when the drugs and diseases are correlated. This might be due to the fact that the methods developed to exploit an underlying group structure benefit from groups containing > 1 causal covariate. Overall, the ARTP performs best, except when there is no overlap between the groups and two of the groups contain all causal covariates. In these cases, the NGL (sum) tends to perform better.

Individual risk prediction

Each method is applied to each of the simulated train data sets. The test sets are subsequently used to assess the prediction performance. Figures A.2.6 – A.2.9 show the box plots with the AUCs, where, in this case, the AUCs reflect how well the methods can predict the ADE occurrence in a patient given his/her drug exposures and disease diagnoses. Figures A.2.6 and A.2.7, on the one hand, show the results when there is *no overlap* between the groups. Figures A.2.8 and A.2.9, on the other hand, show the results when the groups *overlap*. The results when each of the 10 causal covariates being in different groups are shown in Figures A.2.6 and A.2.8. Figures A.2.7 and A.2.9 show the results when the causal covariates are split among only two groups equally. As before, the performance of a random classifier is depicted in each plot with a dashed line.

The figures show that the ARTP performs best when the signal is weak, i.e., OR $=$ 1.5. When the signal is strong, OR $= 5$, the AL shows the best performance. In contrast to inference results, we see that, in all cases, the performance decreases if the covariates are correlated. In particular, in case of OR $= 1.5$ and correlated covariates seems to be the most challenging. The majority of the methods perform only slightly better than a random classifier would do.

Figures and tables

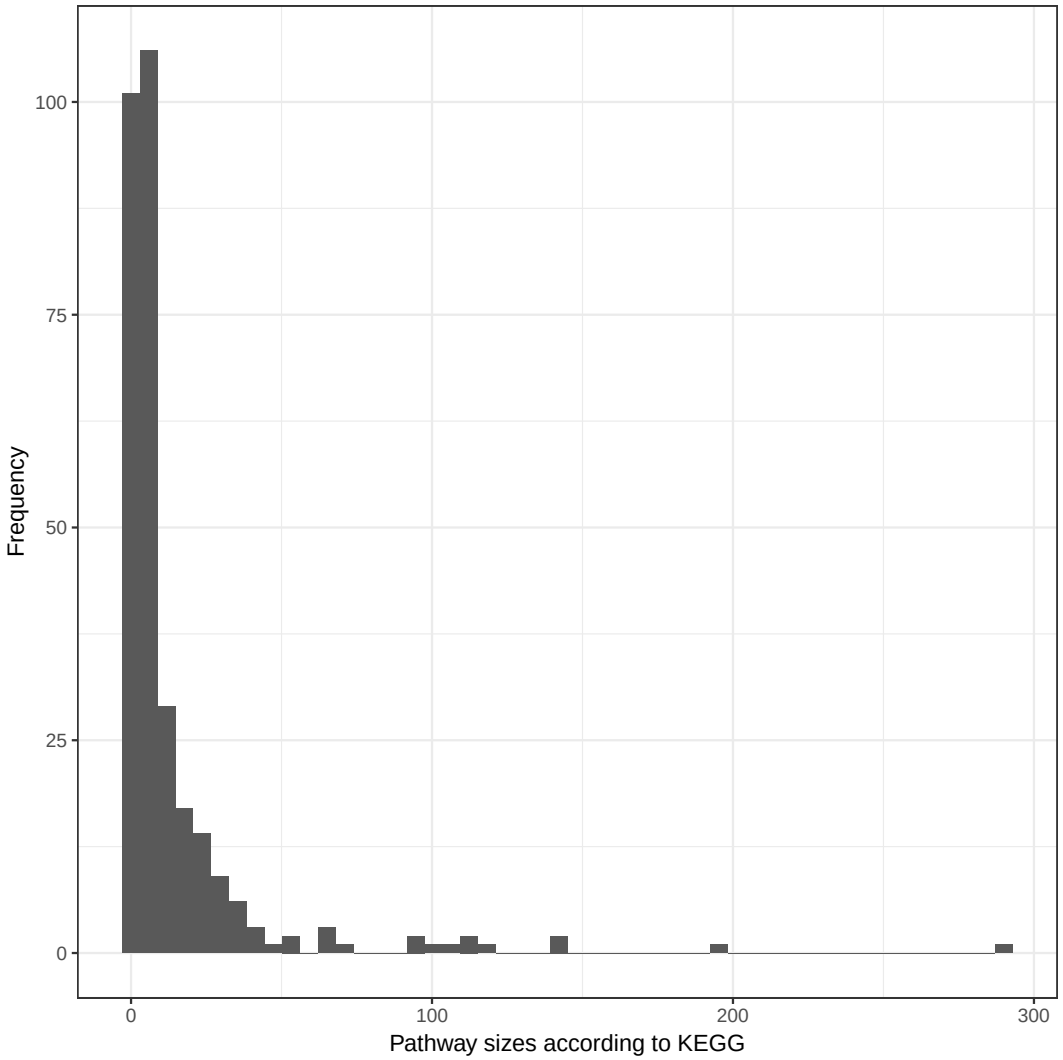


FIGURE A.2.1: Empirical distribution of the number of drugs and diseases in all functional target groups in the KEGG database. The total number of groups is equal to 303. The median group size is 4.

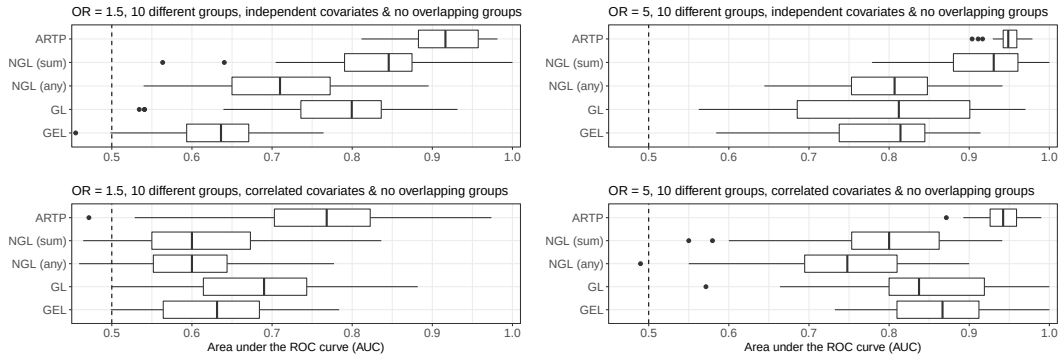


FIGURE A.2.2: Box plots representing AUCs. The AUCs reflect to what extent the methods are able to *infer* which groups have an effect on the ADE risk. Each of the 10 causal covariates are in a different group. There is no overlap between the groups.. The left and right columns show the results when the effect is weak (OR = 1.5) or strong (OR = 5), respectively. The top and bottom rows show the results when the drugs/diseases are independent or correlated, respectively.

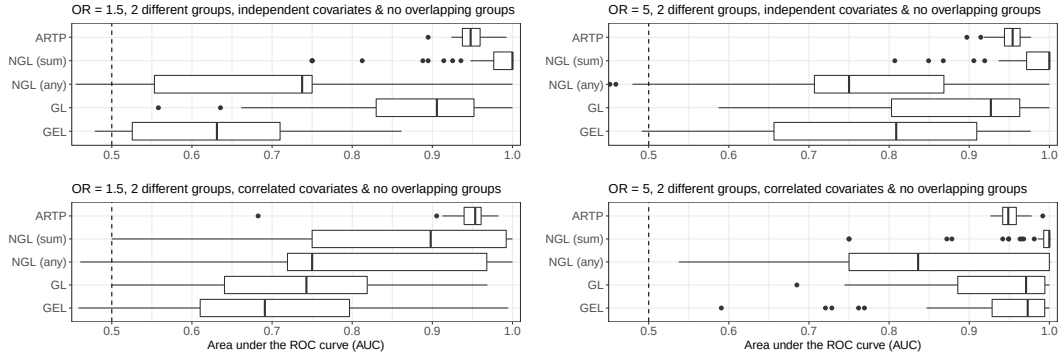


FIGURE A.2.3: Box plots representing AUCs. The AUCs reflect to what extent the methods are able to *infer* which groups have an effect on the ADE risk. Five of the 10 causal covariates are assigned to one group. The other five are assigned to a different group. There is no overlap between the groups.. The left and right columns show the results when the effect is weak (OR = 1.5) or strong (OR = 5), respectively. The top and bottom rows show the results when the drugs/diseases are independent or correlated, respectively.

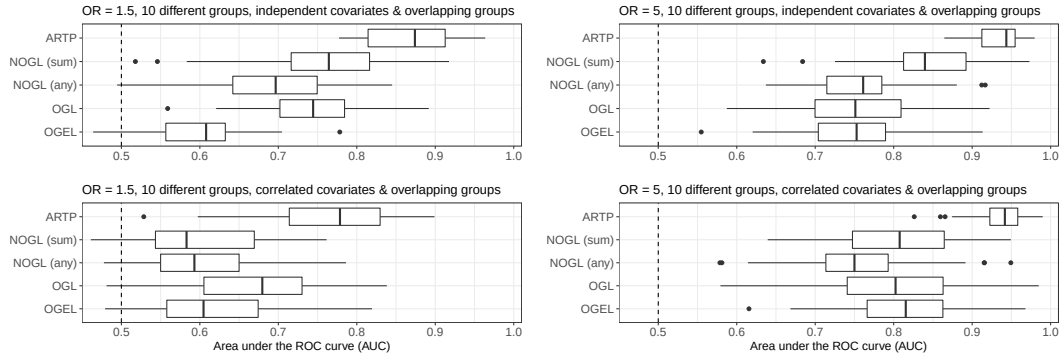


FIGURE A.2.4: Box plots representing AUCs. The AUCs reflect to what extent the methods are able to *infer* which groups have an effect on the ADE risk. Each of the 10 causal covariates are in a different group. Note that due to that fact that the groups, some of the causal covariates might be in another group as well.. The left and right columns show the results when the effect is weak (OR = 1.5) or strong (OR = 5), respectively. The top and bottom rows show the results when the drugs/diseases are independent or correlated, respectively.

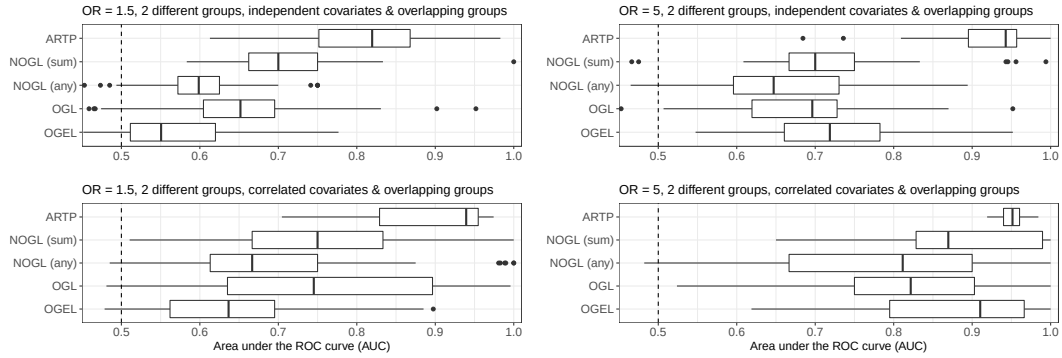


FIGURE A.2.5: Box plots representing AUCs. The AUCs reflect to what extent the methods are able to *infer* which groups have an effect on the ADE risk. Five of the 10 causal covariates are assigned to one group. The other five are assigned to a different group. Note that due to that fact that the groups, some of the causal covariates might be in another group as well.. The left and right columns show the results when the effect is weak (OR = 1.5) or strong (OR = 5), respectively. The top and bottom rows show the results when the drugs/diseases are independent or correlated, respectively.

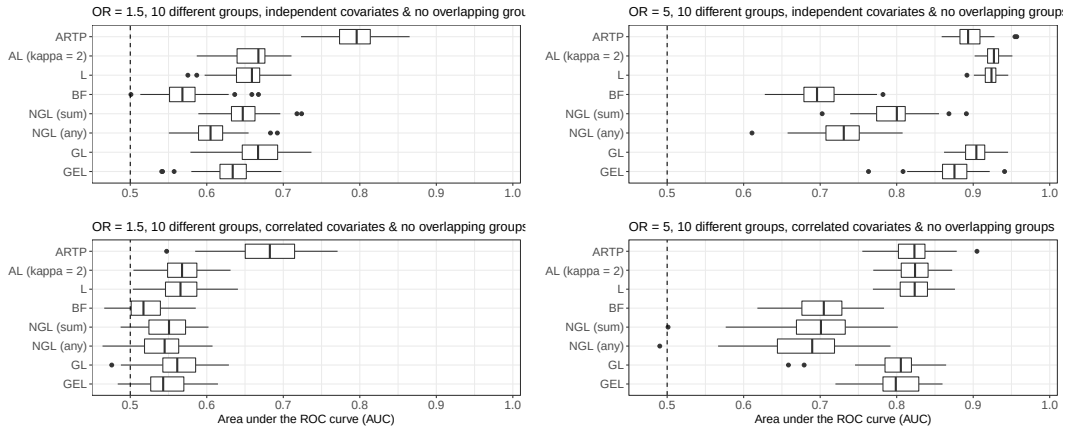


FIGURE A.2.6: Box plots representing AUCs. The AUCs reflect to what extent these methods are able to *predict* whether or not a patient will experience the ADE given his/her drug exposures and diseases. Each of the 10 causal covariates are in a different group. There is no overlap between the groups.. The left and right columns show the results when the effect is weak (OR = 1.5) or strong (OR = 5), respectively. The top and bottom rows show the results when the drugs/diseases are independent or correlated, respectively.

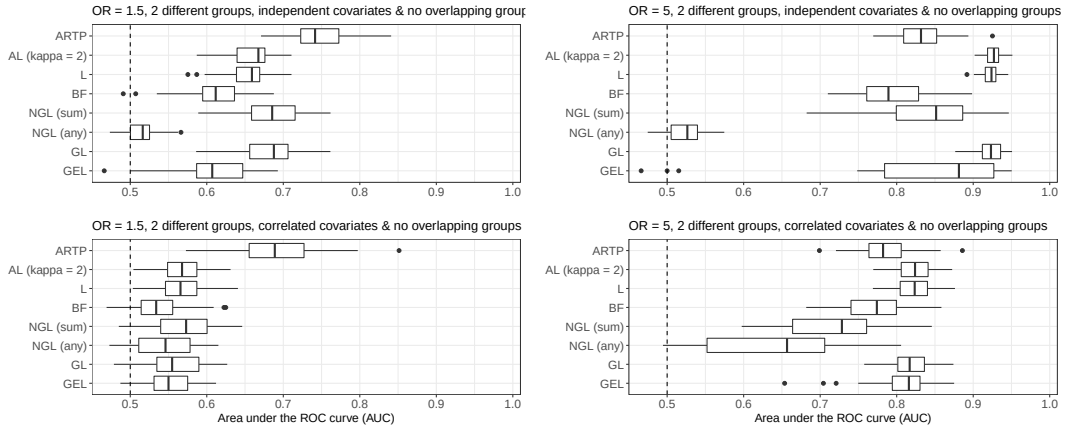


FIGURE A.2.7: Box plots representing AUCs. The AUCs reflect to what extent these methods are able to *predict* whether or not a patient will experience the ADE given his/her drug exposures and diseases. Five of the 10 causal covariates are assigned to one group. The other five are assigned to a different group. There is no overlap between the groups.. The left and right columns show the results when the effect is weak (OR = 1.5) or strong (OR = 5), respectively. The top and bottom rows show the results when the drugs/diseases are independent or correlated, respectively.

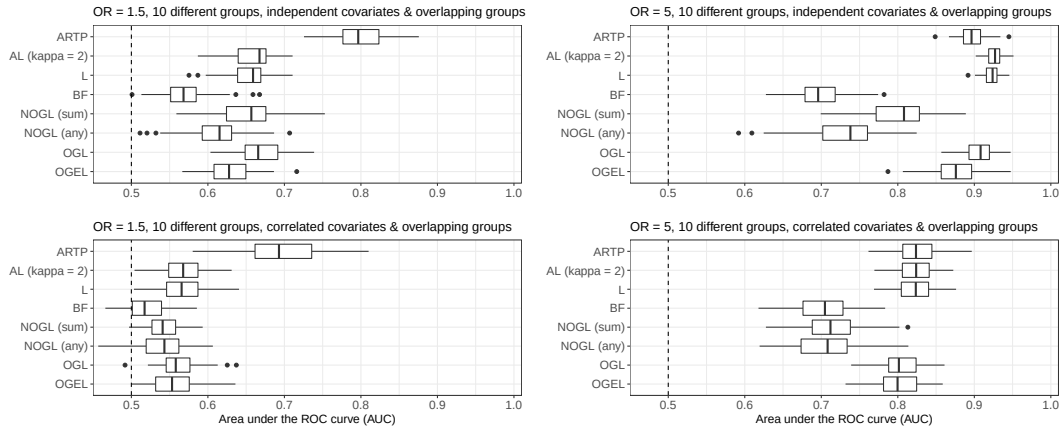


FIGURE A.2.8: Box plots representing AUCs. The AUCs reflect to what extent these methods are able to *predict* whether or not a patient will experience the ADE given his/her drug exposures and diseases. Each of the 10 causal covariates are in a different group. Note that due to that fact that the groups, some of the causal covariates might be in another group as well.. The left and right columns show the results when the effect is weak (OR = 1.5) or strong (OR = 5), respectively. The top and bottom rows show the results when the drugs/diseases are independent or correlated, respectively.

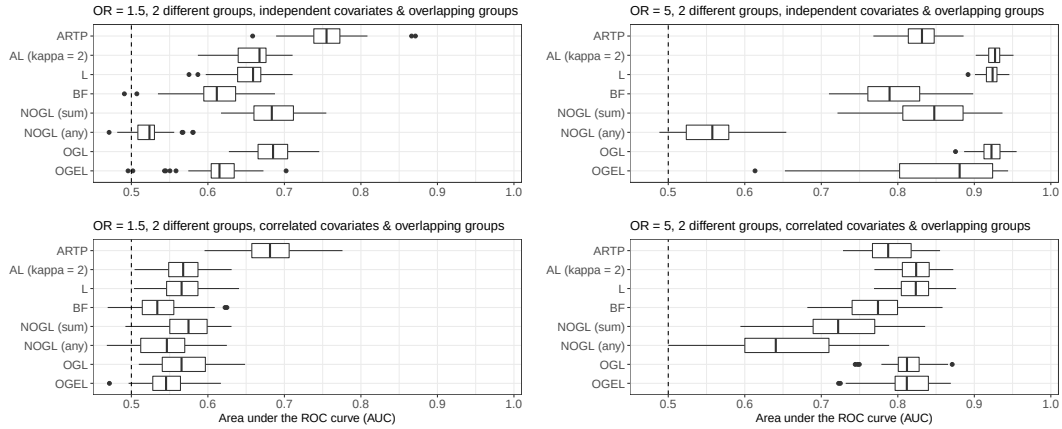


FIGURE A.2.9: Box plots representing AUCs. The AUCs reflect to what extent these methods are able to *predict* whether or not a patient will experience the ADE given his/her drug exposures and diseases. Five of the 10 causal covariates are assigned to one group. The other five are assigned to a different group. Note that due to that fact that the groups, some of the causal covariates might be in another group as well.. The left and right columns show the results when the effect is weak (OR = 1.5) or strong (OR = 5), respectively. The top and bottom rows show the results when the drugs/diseases are independent or correlated, respectively.

TABLE A.2.1: Simulation study statistical methods and the used penalties. Each method is abbreviated, see column 'Abbr.'. All methods are used for risk prediction (see column 'P'), some for group inference (see column 'I').

Abbr.	**Method**	**Penalty $\mathbf{P}(\beta \mid \boldsymbol{\lambda})$**	**P**	**I**	**References**
ARTP	Adaptive rank truncated product†	–	✓	✓	(Yu et al., 2009)
BF	Block forest	–	✓	–	(Hornung and Wright, 2019)
L	Lasso	$\lambda\|\beta\|_1$	✓	-	(Tibshirani, 1996)
AL	Adaptive lasso	$\lambda \sum_{j=1}^{m} \|\tilde{\beta}\|_1^{-\kappa} \|\beta_j\|_1$	✓	-	(Zou, 2006)
NGL	Naïve group lasso	$\lambda\|\beta\|_1$	✓	✓	–
NOGL	Naïve overlapping* group lasso	$\lambda\|\gamma^{(g)}\|_1$	✓	✓	–
GL	Group lasso	$\lambda \sum_{g=1}^{G} \|\sqrt{s_g}\beta^{(g)}\|_2$	✓	✓	(Meier et al., 2008)
OGL	Overlapping* group lasso	$\lambda \sum_{g=1}^{G} \|\sqrt{s_g}\gamma^{(g)}\|_2$	✓	✓	(Zeng and Breheny, 2016)
GEL	Group exponential lasso	$\sum_{g=1}^{G} \frac{\lambda^2}{\tau} \left(1 - \exp\left\{-\frac{\tau\|\beta\|_1}{\lambda}\right\}\right)$	✓	✓	(Breheny, 2015)
OGEL	Overlapping* group exponential lasso	$\sum_{g=1}^{G} \frac{\lambda^2}{\tau} \left(1 - \exp\left\{-\frac{\tau\|\gamma^{(g)}\|_1}{\lambda}\right\}\right)$	✓	✓	(Breheny, 2015)

λ: tuning parameter; β: regression coefficient vector; $\tilde{\beta}$: pilot estimate vector; s_g: size of group g; $\gamma^{(g)}$ latent coefficient vector satisfying $\sum_{g=1}^{G} \gamma^{(g)} = \beta$
$\kappa \in \{.5, 1, 1.5, 2\}$; $\tau = 1/3$
*overlapping regression types use transformed design matrix with duplicated columns for group overlapping covariates
†the set $\{1, 2, \ldots, 5\}$ is used as candidate truncation points

TABLE A.2.2: Simulation study parameter settings.

Description	**Notation**	**Values**
total number of patients	n	2,000
number of patients in train set	–	1320
number of patients in test set	–	680
total number of drugs/diseases	m	1,000
number of causal covariates	m_a	10
number of causal covariates per group	–	1 or 5
number of repetitions	–	50
probability of experiencing the ADE	–	50%
odds ratio between causal covariate and ADE	OR	1.5, 3, or 5
correlation between covariates	ρ	0 or .25

A.3 Integration and modeling of secondary health data: Data dimensions and statistics

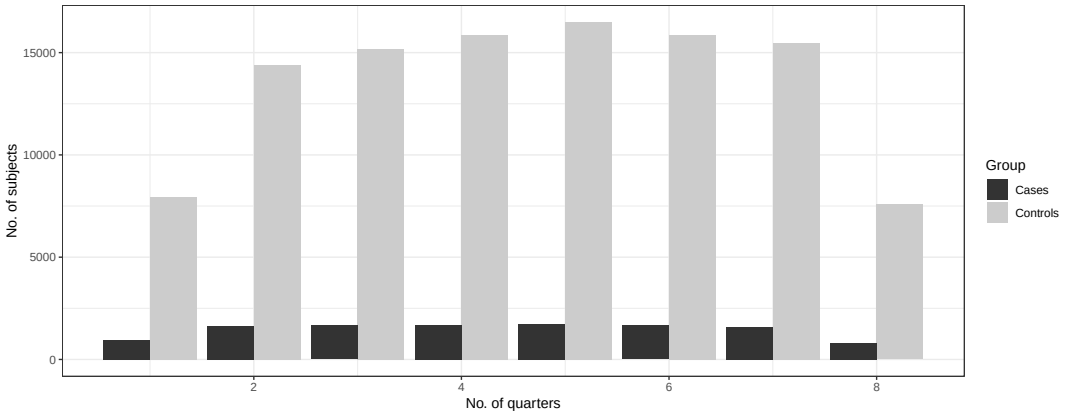


FIGURE A.3.10: A bar plot of patient time in calendar quarters in the case-control data set. It is showing a near normal distribution.

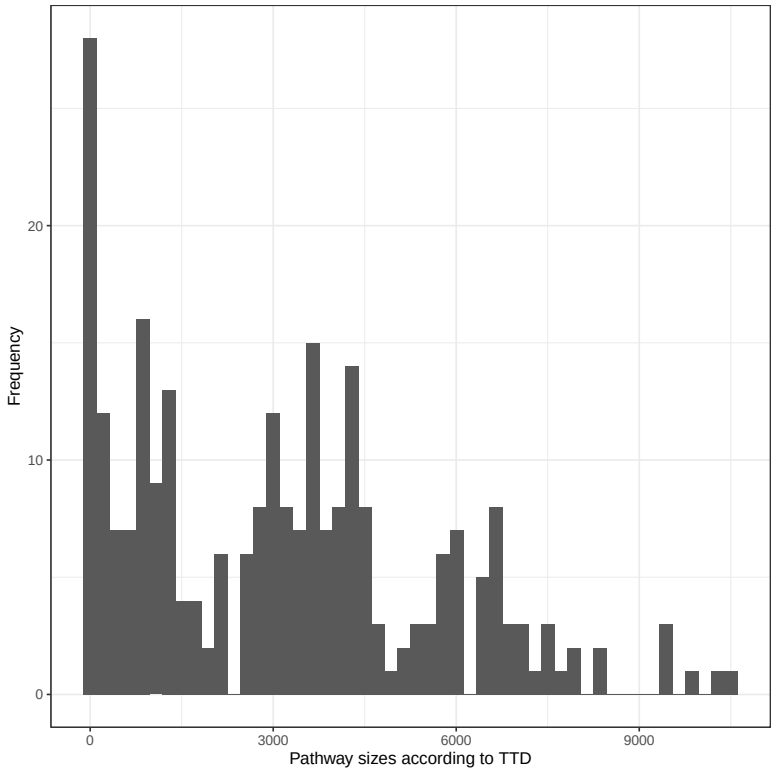


FIGURE A.3.11: The empirical distribution of the number of drugs and diseases in all functional target groups in the curated TTD data set.

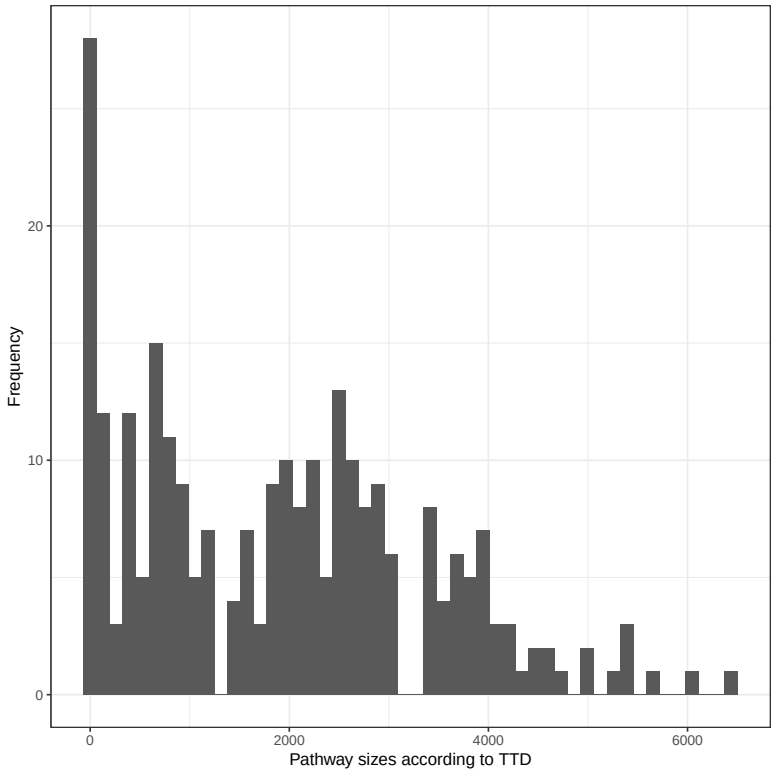


FIGURE A.3.12: The empirical distribution of the number of non-zero variance drugs and diseases in all functional target groups in the GePaRD data set according to TTD grouping.

TABLE A.3.3: Number of rows, unique covariates per patient (i.e., first incidence) and number of covariates in eligible subjects data in GePaRD.

Module	No. rows	Unique covariates per subject	No. covariates
Inpatient diagnosis	22,609,126	14,131,149	11,329
Outpatient diagnosis	497,891,376	171,436,672	13,965
Dispensation	104,057,634	36,522,387	2,410

TABLE A.3.4: Descriptive statistics of the matched case-control data set with respect to socio-demographics and data dimensions. The non-vitamin K oral anticoagulants (NOACs) considered are rivaroxaban, apixaban, edoxaban, and dabigatran. Numbers of covariates include zero variance predictors.

	Cases (n = 11,717)	**Controls** (n = 108,747)	**Total** (n = 120,464)
Age (in years; mean $\pm$ SD)	59.01 ± 17.92	59.55 ± 17.83	59.5 ± 17.84
Sex (no.; % female)	5,534; 47.2%	52,580; 48.35%	58,114; 48%
NOAC users (no.; %)	619; 5.28%	3,238; 2.97%	3,857; 3.2%
Max no. drugs per subject	69	54	1,711
Max no. diseases per subject	301	198	10,693

TABLE A.3.5: Performance of statistical methods measured as recall, precision and F1-score. L = LASSO, OGL = overlapping group LASSO, ARTP = adaptive combination of rank truncated product, BF = block forests, NGL = naive group LASSO, and SM = standard model for logistic regression.

	L	**OGL**	**ARTP**	**BF**	**NGL**	**SM**
Recall	0.28	0	0	0.73	0.15	0.35
Precision	0.72	0	0	0.074	0.54	0.67
F1-score	0.4	0	0	0.13	0.24	0.46

Appendix B

Publications

B.1 Deciphering patterns of adaptation and acclimation in the transcriptome of *Phaeocystis antarctica* to changing iron conditions

Contribution to the manuscript: I maintained and inoculated the cultures, harvested the cells and extracted RNA with S. Beszteri. I have executed the transcriptome assembly, analysis and differential expression inference with Harms L supervised by S. Frickenhaus and A. Moustafa. I participated in the conceptualization of the manuscript, and I wrote the initial draft of the manuscript with S. Frickenhaus and S. Beszteri. I processed the sequence-based data to be deposited and publicly available through NCBI. The work was done in collaboration with and under the supervision of the co-authors. The manuscript is published in the Journal of Phycology.

J. Phycol. **56,** 747–760 (2020)

DOI: 10.1111/jpy.12979

DECIPHERING PATTERNS OF ADAPTATION AND ACCLIMATION IN THE TRANSCRIPTOME OF *PHAEOCYSTIS ANTARCTICA* TO CHANGING IRON CONDITIONS[1]

Mariam R. Rizkallah[2]

Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany

Stephan Frickenhaus

Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany
Centre for Industrial Mathematics, University of Bremen, Bibliothekstrasse 1, 28359 Postfach 330440, 28334 Bremen, Germany

Scarlett Trimborn

Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany
Department of Marine Botany, University of Bremen, Bibliothekstrasse 1, 28359 Postfach 330440, 28334 Bremen, Germany

Lars Harms

Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany
Helmholtz Institute for Functional Marine Biodiversity at the University of Oldenburg (HIFMB), Ammerländer Herrstrasse 231, 26129 Oldenburg, Germany

Ahmed Moustafa

Department of Biology, American University in Cairo, P.O. Box 74, 11835 Cairo, Egypt

Vladimir Benes

European Molecular Biology Laboratory, Meyerhofstraße 1, 69117 Heidelberg, Germany

Steffi Gäbler-Schwarz and Sara Beszteri[2]*

Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany

The haptophyte *Phaeocystis antarctica* is endemic to the Southern Ocean, where iron supply is sporadic and its availability limits primary production. In iron fertilization experiments, *P. antarctica* showed a prompt and steady increase in cell abundance compared to heavily silicified diatoms along with enhanced colony formation. Here we utilized a transcriptomic approach to investigate molecular responses to alleviation of iron limitation in *P. antarctica*. We analyzed the transcriptomic response before and after (14 h, 24 h and 72 h) iron addition to a low-iron acclimated culture. After iron addition, we observed indicators of a quick reorganization of cellular energetics, from carbohydrate catabolism and mitochondrial energy production to anabolism. In addition to typical substitution responses from an iron-economic toward an iron-sufficient state for flavodoxin (ferredoxin) and plastocyanin (cytochrome c_6), we found other genes utilizing the same strategy involved in nitrogen assimilation and fatty acid desaturation. Our results shed light on a number of adaptive mechanisms that *P. antarctica* uses under low iron, including the utilization of a Cu-dependent ferric reductase system and indication of mixotrophic growth. The gene expression patterns underpin *P. antarctica* as a quick responder to iron addition.

Key index words: **Antarctic regions; haptophyta; iron; photosynthesis; phytoplankton; transcriptome**

***Abbreviations*: DEG, differentially expressed gene; HNLC, high-nitrate low-chlorophyll; PPP, pentose phosphate pathway; SO, Southern Ocean; TCA cycle, tricarboxylic acid cycle**

The Southern Ocean (SO) is the largest high-nitrate low-chlorophyll (HNLC) region with sub-nanomolar concentrations of total dissolved iron, abundant concentrations of macronutrients yet low

[1]Received 22 January 2019. Accepted 21 January 2020. First Published Online 18 February 2020. Published Online 15 April 2020, Wiley Online Library (wileyonlinelibrary.com).
[2]Current address: Department of Biodiversity, University of Duisberg-Essen, 45117 Essen, Germany.
[2]Authors for correspondence: e-mail m.rizkallah@jacobs-university.de; sara.beszteri@uni-due.de.
Editorial Responsibility: T. Mock (Associate Editor)

rates of nitrate uptake, and dominance of pico- and nanophytoplankton species (Dugdale and Wilkerson 1991, Smetacek et al. 1997, Assmy et al. 2007, Marchetti et al. 2012, Trimborn et al. 2017). Iron supply to the SO includes dust deposition and melting icebergs (Assmy et al. 2007, Boyd et al. 2012), but as iron remains bound to organic ligands and therefore biologically unavailable to phytoplankton (Maldonado et al. 2005, Shaked and Lis 2012, Groussman et al. 2015, Hutchins and Boyd 2016), it is limiting phytoplankton growth and productivity (Martin et al. 1990).

Iron is essential for phytoplankton growth as it serves as an electron carrier in photosynthesis and mitochondrial respiration. It is also required as a cofactor in countless processes such as fatty acid biosynthesis, nitrate reduction and assimilation (Marchetti et al. 2012, Harel et al. 2014, Schoffman et al. 2016). The effect of iron limitation has only been studied in temperate diatoms at the molecular level (Allen et al. 2008, Lommer et al. 2012, Morrissey and Bowler 2012, Raven 2013, Smith et al. 2016), showing how the newly acquired iron is allocated (Strzepek and Harrison 2004, Lommer et al. 2012, Marchetti et al. 2012, Smith et al. 2016). Haptophytes as well showed similar adaptation to iron limitation, and lower iron requirements for growth (Strzepek et al. 2011, 2012, 2019).

Phaeocystis is a cosmopolitan genus within the division of haptophytes. Its three colony- and bloom-forming species are: the temperate *P. globosa* in the North Sea, the Arctic *P. pouchetii* and the Antarctic *P. antarctica* (Schoemann et al. 2005, Verity et al. 2007, Beardall et al. 2009). Colonial life stage provides these species with protective and competitive advantages over the solitary stage, with the protein–carbohydrate colony skin serving as a mechanical barrier against infections, and the large colony size protecting against grazers (Hamm 2000). The mucilaginous structure of the colony matrix further allows for storage of micro- (iron and manganese) and macro- (carbon and nitrogen) nutrients (Hamm 2000, Schoemann et al. 2005, Gaebler-Schwarz et al. 2010).

Phaeocystis antarctica is endemic to the largely iron-limited SO and forms large blooms, which are frequently recorded in the iron-enriched shelf areas such as Ross Sea and Prydz Bay (Schoemann et al. 2005, Smith et al. 2014b). In vitro experiments showed that *P. antarctica* has a strong response to iron limitation as indicated by reduction in its growth rates and photosynthetic fitness (Strzepek et al. 2011, Alderkamp et al. 2012), whereas iron addition was reported to increase growth rates and trigger colony formation in *P. antarctica* (Bender et al. 2018). In situ iron fertilization experiments in the SO reported haptophytes (*P. antarctica)* among the groups contributing to the elevation in chlorophyll *a* signal after iron enrichment (Gall et al. 2001, Boyd 2002b, de Baar et al. 2005). In particular, in the iron fertilization experiment EisenEx, *P. antarctica* showed a prompt and steady increase in cell abundance compared to heavily silicified diatoms, in addition to a higher frequency of colony formation (Assmy et al. 2007). In the subarctic Pacific, metatranscriptomics showed that haptophytes (*P. globosa*) utilized added iron faster than diatoms, with an overexpression of photosynthesis genes (Marchetti et al. 2012).

Adaptation and acclimation are considered types of stress response (Borowitzka 2018). Acclimation, on the one hand, is the change in phenotype (through changes in gene expression) in response to stress in an attempt to restore homeostasis in the cell. Once acclimation is accomplished and homeostasis is restored, the cells are no longer considered stressed. Adaptation, on the other hand, is the change in the genotype of the organism in response to environmental changes. In other words, adaptation can engrave the acclimated phenotype in the cell's genome after the necessary number of generations has been successfully acclimated to the stressful conditions (Borowitzka 2018). *Phaeocystis antarctica* was found to combat photosynthesis-limiting factors such as low iron by increasing photosynthetic iron use efficiencies (e.g., replacement of iron rich with iron-economic photosynthetic components; Strzepek et al. 2019).

Here, we aim at deciphering the molecular basis of adaptation to low iron availability and its subsequent acclimation following iron enrichment in the ecologically relevant prymnesiophyte *Phaeocystis antarctica,* a colony-forming species isolated from the Ross Sea. We highlight the molecular processes that might be the basis of the adaptation of *P. antarctica* to iron limitation, and its acclimation to iron addition. We provide novel evidence based on gene expression data that supports possible mixotrophic behavior of *P. antarctica* cells under iron limitation.

MATERIALS AND METHODS

Culture conditions. A colony-forming strain of *Phaeocystis antarctica* (strain #25 isolated from Ross Sea [76° S; 170° W] in 2003) was acclimated in f/2 growth medium where iron was omitted from the trace metal mix (pH 8.0–8.3; Guillard and Ryther 1962) prepared with Southern Ocean seawater. Iron-free trace metal mixture and desferrioxamine B (DFB) chelator (10 nmol $\cdot$ L^{-1} final concentration; Strzepek et al. 2011) were syringe filtered through cellulose acetate 0.22 μm sterile filters (Cole-Parmer, Montreal, Canada) before its addition to the natural seawater. The iron-limited culture was used to inoculate quadruplicates with starting cell concentration of 2×10^4 cells $\cdot$ mL^{-1}. Cultures were acclimated at 2°C under a 16:8 h light:-dark cycle (40 μmol photons $\cdot$ m^{-2} $\cdot$ s^{-1}; Philips Master TL-D 18 W daylight lamps with neutral density screens with the lights switched on at 6 AM and off at 10 PM). All cultures were incubated in 2L polycarbonate bottles (Nalgene, New York, USA) which were detergent and acid treated (3-day 0.1% CITRANOX®-bath followed by 7-d 0.1N HCl-bath) and rinsed with ultrapure Milli-Q® water (Millipore, Darmstadt, Germany). The cell densities and physiologic status of the quadruplicates are summarized in Table S1 in the Supporting Information.

Forty-eight hours after inoculation, iron-deplete quadruplicates were supplemented with 1800 nM of iron (as $FeCl_3.6H_2O$ dissolved in ultrapure Milli-Q® water, syringe-filtered). The

main experiment was conducted for 5 d. Cells were harvested from the iron-limited and iron-enriched incubation bottles at 0 h, 14 h, 24 h and 72 h and 100 ± 10 mL of each culture was filtered through MF-Millipore™ membrane filters (1.2 μm; Merck KGaA, Darmstadt, Germany) using vacuum filtration. Cells were resuspended in 500 μL beta-mercaptoethanol / RLT buffer (Qiagen, Hilden, Germany) and preserved in liquid N_2 at −80°C until RNA extraction. Samples were taken in the evening (6 PM; 12 h of light) except for 14 h (10 AM; 4 h of light). Information on the physiologic assessments, and iron and light status of treatment cultures is provided in Table S2 in the Supporting Information.

RNA extraction, qualitative and quantitative analysis, and sequencing. Total RNA was extracted using RNeasy® Plant Mini Kit (Qiagen) as published before (Beszteri et al. 2012). RNA concentration ($ng \cdot L^{-1}$) was estimated using a NanoDrop® ND-1000 spectrophotometer (Peqlab, Erlangen, Germany). RNA integrity (RIN) was estimated using a 2100 Bioanalyzer coupled with 2100 Expert Software (Agilent Technologies Inc., Boeblingen, Germany). RNA (260/280 > 1.6 and RIN > 5) was processed by The European Molecular Biology Laboratory (EMBL) Genomic Core Facilities (GeneCore, EMBL Heidelberg, Germany) for complementary DNA (cDNA) library construction of poly(A) RNA, and for paired-end RNA sequencing using Illumina HiSeq2000 sequencer (Illumina Inc., San Diego, CA, USA).

De novo transcriptome assembly and functional analysis. Trimmomatic (v0.32; Bolger et al. 2014) was used to trim sequencing adapters and to eliminate bases of Phred quality scores below 15 and reads shorter than 30 bases. Quality-filtered paired-end reads from the iron-limited culture that was used to inoculate quadruplicates, the iron-deplete control and iron-replete treatment were used for assembly using Trinity de novo transcriptome assembler pipeline (v2.0.4; Grabherr et al. 2011, Haas et al. 2013). Open reading frames (ORFs) identification and translation were performed using TransDecoder (v2.0.1) accounting for homology search results from UniProtKB/Swiss-Prot (r2015-09) and Pfam-A (r27.0). Translated ORFs were analyzed using Trinotate (v2.0; *e*-value ≤ 1e-5). Translated ORFs were compared to *Phaeocystis antarctica* peptide sequences (iMicrobe sample MMETSP1100; Koid et al. 2014) using OrthoMCL (v2.0.9; Li et al. 2003). Transcriptome functional coverage was estimated by comparing the assembled transcripts against the eukaryotic Benchmarking Universal Single-Copy Orthologs (BUSCO; v1.22; *e*-value ≤ 1e-5; Simão et al. 2015). Assembled transcripts were compared against *P. antarctica* mitochondrial and plastid genomes (Smith et al. 2014a) using nucleotide BLAST mapping cDNA/EST to a genome protocol (stringent reward/penalty, *e*-value ≤ 1e-100 and query coverage ≥ 90). Hits were visualized using BLASTGrabber 2.0 (Neumann et al. 2014) and overlapping hits segments were fused. Project metadata is available at BioProject (Record: PRJNA395466). Quality-filtered raw sequencing reads are available at the Sequence Read Archive (SRA; Study accession: SRP113407). Standard quality-managed data of this Transcriptome Shotgun Assembly project have been deposited at DDBJ/EMBL/GenBank (Accession: GFUQ00000000).

Differential gene expression and functional enrichment. Abundance of the generated transcripts was estimated for each replicate by RNA-Seq by Expectation Maximization (RSEM; Li and Dewey 2011). Sample correlation was assessed through hierarchical clustering of the fragments per feature Kb per million reads mapped (FPKM) values using Pvclust (Suzuki and Shimodaira 2006). Differential expression analysis was conducted comparing each iron-response time point (14 h, 24 h and 72 h) against 0 h by DESeq2 (Love et al. 2014). Only transcript contigs (denoted as "genes" for simplicity) of ≥ 300 bases and sum of rounded counts ≥ 40 were considered for downstream differential expression analysis. Genes with false discovery rate (FDR) ≤ 0.001 and absolute log2 fold-change ≥ 2 were considered differentially expressed. Heatmaps were used for visualizing the hierarchical clustering of normalized expression values of the differentially expressed genes (DEGs) based on Euclidean distance with complete linkage. DEG clusters were identified through manual inspection by *k*-mean clustering of time-point averaged normalized expression (i.e., FPKM), where the number of centers was set to seven.

The number was validated through NbClust (Charrad et al. 2014) and manual inspection. Gene Ontology (GO) enrichment analysis was conducted on the DEG clusters through Trinity (one-sided Fisher's exact test) with significance cut-off of FDR ($N_{annotated}/N_{GO\text{-}annotated}$) ≤ 0.05. Over-represented GO terms were categorized using Categorizer (Na et al. 2014). DEG clusters were also inspected for orthologous group frequencies. The frequencies were derived from Trinotate's mapping of BLAST results to Evolutionary Genealogy of Genes (eggNOG; v4.0). Functional annotation of DEGs was based on best UniProt and Pfam hits of the longest ORF and excluding mammalian hits. Count and normalized gene expression matrices as well as transcript annotations are available at the Gene Expression Omnibus (GEO) database (Accession: GSE102608).

RESULTS

Transcriptome characteristics. To profile the transcriptomic characteristics of *Phaeocystis antarctica*, we sequenced mRNA from the iron-limited culture, and the iron-limited control (0 h) and -enriched treatments (14 h, 24 h and 72 h). A total of 389,846,414 reads were sequenced and quality filtered into 312,273,819 reads prior to assembly. The total number of bases of the final assembly is 87,421,418 assembled into 122,927 transcripts (i.e., isoforms) of 110,971 contigs (i.e., hypothetical genes). Transcripts N_{50} was 953 bp and estimated GC content was 63.16% (Table 1). The frequency distribution of isoforms mapping showed that the largest fraction of genes (94%) constituted of unique transcripts

TABLE 1. *Phaeocystis antarctica* transcriptome statistics. Gene length is the length of the longest transcript (i.e., isoform) of the gene.

Category	Number	Total bases	N50 (bp)	Mean (bp)	Median (bp)
Sequenced reads	389,846,414	19,882,167,114	–	–	–
Post QC reads	345,183,306	17,179,127,605	–	–	–
Assembled reads	312,273,819	15,925,964,769	–	–	–
Total genes	110,971	76,380,968	916	688.3	486
Total transcripts	122,927	87,421,418	953	711.17	506

(Fig. S1 in the Supporting Information). To assess sample correlation, we clustered hierarchically the raw counts of the assembled hypothetical genes (Fig. S2 in the Supporting Information).

A total of 105,163 open reading frames (ORFs) were predicted and translated. Transcripts were annotated based on similarity search against UniProt database and domain search against Pfam database. Annotations were cross-referenced with GO and eggNOG (Fig. 1; largest gene families are provided in Table S3 in the Supporting Information). Here for consistency, we report the results at the gene level (detailed statistics are given in Table 2).

We assessed the evolutionary origin of the assembled genes. Generally, 16,419 genes (14.8% of total transcriptome genes) were of eukaryotic origin excluding mammals. Specifically, 6,455 genes (5.8%) were closest to Streptophyta, 337 (0.3%) to Chlorophyta and 92 (0.08%) to Haptophyta. There were 3,878 (3.5%) genes of bacterial origin, of which 459 genes (0.41%) were of cyanobacterial origin. The number of genes of archaeal and viral origins accounted 250 (0.22%) and 206 (0.19%), respectively.

We compared the study transcriptome against the published *Phaeocystis antarctica* transcriptome from the MMETSP project (Koid et al. 2014; MMETSP1100, a data set that contains 53,204 coding nucleotide sequences and 54,300 translated peptide sequences) in terms of sequence overlap using OrthoMCL (Li et al. 2003). About 9% of the study unique (i.e., deduplicated) translated ORFs were orthologs of 25% of MMETSP's *P. antarctica* coding sequences. Furthermore, we compared the transcriptomes in terms of functional coverage against BUSCO's eukaryotic gene set (429 orthologs; Table 3). Additionally, we also compared sequence coverage of both transcriptomes against published partial mitochondrial and complete plastid genomes (Smith et al. 2014a; Table 3).

Global patterns of differentially expressed genes. To assess the immediate (14 h), short- (24 h) and long-term (72 h) response following iron enrichment, we compared *Phaeocystis antarctica* gene expression at the different time points after iron addition relative to the one before iron addition (0 h). We found in total 16,895 differentially expressed genes (DEGs) following iron enrichment at the different sampling points: 12,081 genes after 14 h (4,130 up-regulated/7,951 down-regulated), 84,16 genes after 24 h (1,326/7,090), and 5,306 genes after 72 h (1,382/3,924). Time point 14 h differed from the other sampling points with regard to light duration with ~4 h of light at 14 h (Table S2). To focus on the temporal effect of iron enrichment on gene expression, we excluded the DEGs that were only differentially expressed at 14 h from downstream analyses, resulting in 10,715 DEGs (9.7% of total assembled genes).

DEGs were divided into seven clusters based on *k*-means clustering (Fig. S3 in the Supporting Information), to support and explain the functional

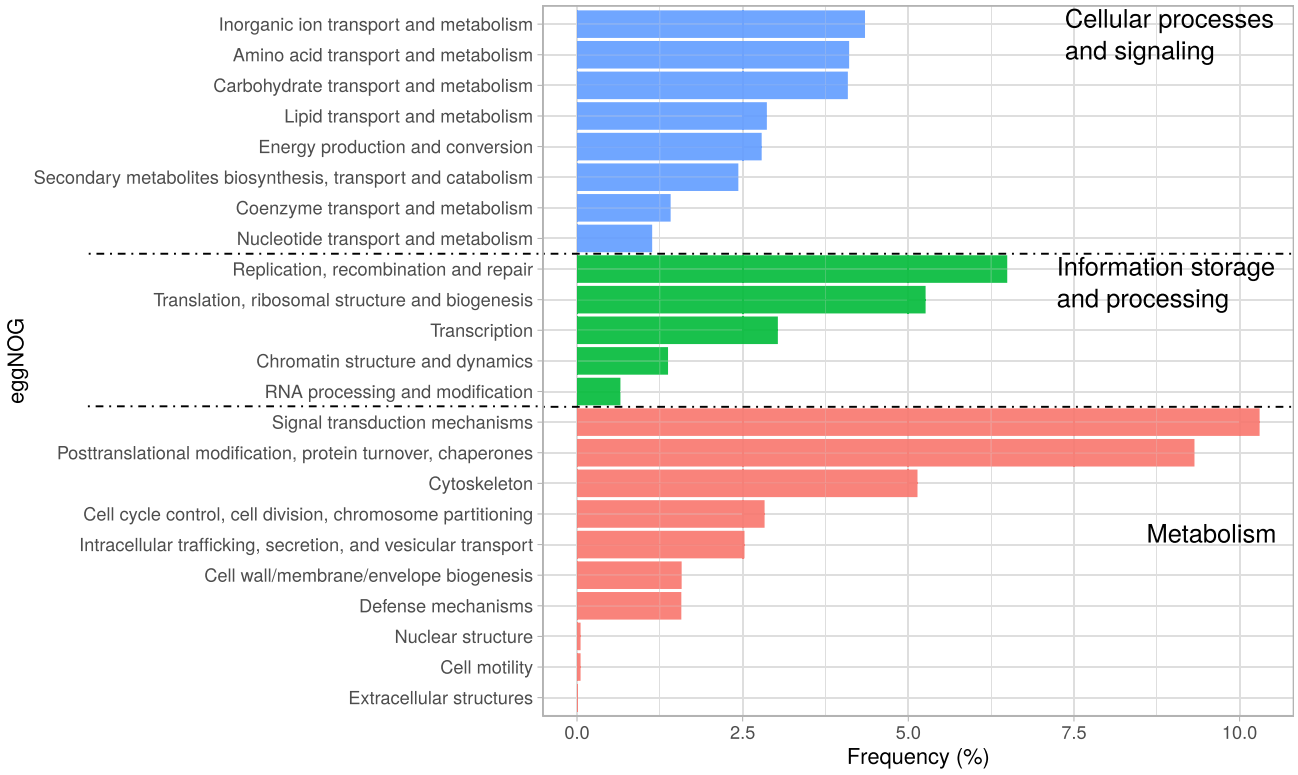


FIG. 1. Gene families assigned to eggNOG orthologous groups in *Phaeocystis antarctica* transcripts. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2. *Phaeocystis antarctica* transcriptome functional annotation statistics in each database.

Database	Gene hits (no.; %)	Transcript hits (no.; %)	ORF hits (no.; %)
UniProt	28,781; 26%	32,134; 26.14%	37,241; 35.4%
Pfam	31,740; 28.6%	35,664; 29%	36,295; 34.5%
GO	26,826; 24.17%	29,807; 24.25%	27,202; 25.9%
eggNOG	16,958; 15.28%	18,683; 15.2%	16,908; 16.07%

TABLE 3. Functional and sequence coverage of nonnuclear genomes of the study and MMETSP *Phaeocystis antarctica* transcriptomes.

	BUSCO Gene Set % (% single-copies; % duplicated)		Organelle Genome (no. nonoverlapping transcripts; % genome length)	
Data set	Complete	Fragmented	Plastid	Mitochondria
Current	62% (41.7%; 20%)	15%	51; 93.4%	17; 73.7%
MMETSP	53% (52.2%; 0.7%)	14.40%	1; 1.32%	0; 0%

analysis below. Clusters A, B, D, and E showed similar patterns of immediate down-regulation and were grouped into cluster 1. The variability within cluster 1 seems to be the largest perhaps due to the fusion of the clusters. We renumbered the clusters C to 2, F to 3, and G to 4 in downstream analysis (Fig. 2). In clusters 1 and 4, a global pattern of down-regulation following iron addition was prevalent, while clusters 2 and 3 showed a pattern of up-regulation. We denote clusters 1 and 4, respectively, "immediate-down" and "delayed-down," and clusters 2 and 3, respectively, "immediate-up" and "delayed or progressive-up" (detailed statistics in Table 4).

Functional analysis of DEGs. GO enrichment and eggNOG statistics: To a total of 3,638 DEGs, we were able to append function based on similarity (according to UniProt and Pfam annotation). Table A in Appendix S1 in the Supporting Information summarizes the results of the GO enrichment analysis of the DEGs. In down-regulation cluster 1 (immediate-down), cell cycle, cell motility, intracellular localization and transport, endo/exocytosis, ATP catabolic process, and tricarboxylic acid cycle, glucan biosynthesis, signaling and calcium transport were overrepresented. In down-regulation cluster 4 (delayed-down), nucleus (i.e., nucleolus and splicing) GO terms were overrepresented. In up-regulation cluster 2 (immediate-up), mitochondria, respiration and translation GO terms, whereas in up-regulation cluster 3, oxidation–reduction processes and signaling GO terms were more abundant. Photosynthesis GO terms were represented in all clusters except cluster 1, while cell motility terms were only present in cluster 1. The frequencies of eggNOG orthologous groups across the clusters were analyzed (Fig. 3; Table B in Appendix S1), and were in agreement with GO enrichment analysis.

Iron acquisition and homeostasis: Altogether, 31 gene candidates implicated in iron assimilation and transport exhibited differential expression (Table C in Appendix S1). After iron addition, 19 of these were immediately down-regulated, including components of a high-affinity iron uptake system (ferric reduction oxidases and multicopper oxidases). Also, ferrochelatase and other genes bearing domains found in iron-starvation-induced proteins ISIP2A and ISIP3 were expressed at the later time points at lower levels compared to 0 h. A potential mitochondrial iron transporter (mitoferrin) also showed sustained down-regulation after iron addition. Contrastingly, NADH-cytochrome b_5 reductases and cytochrome b_5 were immediately up-regulated following iron addition, whereas the two vacuolar iron transporter fragments (VIT11 and VIT1) showed immediate and delayed up-regulation, respectively. Ferritin was not differentially expressed.

Photosynthesis and pigment biosynthesis: A relatively large number of genes related to photosynthesis (46) were differentially expressed under the different time points (Table D in Appendix S1). After iron addition, 38 genes were up-regulated. Several genes of photosystems I and II and of the plastidic electron transport chain (e.g., cytochrome b_6f subunits) were more abundant, whereas flavodoxin and plastocyanin were down-regulated immediately after iron addition.

Light-harvesting complex genes and photoreceptors (67; Table E in Appendix S1) exhibited differential expression according to their function: those implicated in light harvesting were more abundant, whereas others involved in photoprotection became down-regulated in response to iron addition. In contrast, the genes coding for the chlorophyll *a/b* binding protein L1818 were immediately up-regulated after iron addition. Furthermore, several genes involved in chlorophyll and accessory pigment biosynthesis (26; Table F in Appendix S1) were mostly more abundant after iron addition. Chlorophyll, xanthophyll and carotenoid biosynthesis genes were mainly differentially expressed at the light (i.e., morning) time point at 14 h.

Nitrogen and sulfur assimilation and metabolism: Table G (Appendix S1) depicts the DEGs related to nitrogen and sulfur metabolism (26 genes). Strikingly, four of six nitrite reductases (NiR) were down-regulated after iron addition, whereas the other two fragments exhibited higher expression. Similarly, two of three sulfite reductases and three of four glutamate synthases (GS) were down-regulated after iron addition, whereas one sulfite reductase and one GS were up-regulated. Also, type-3 glutamine synthases showed similar mixed expression patterns, whereas ammonium transporters were mostly immediately down-regulated after iron enrichment. Interestingly, all potential nitrate reductases were immediately up-regulated.

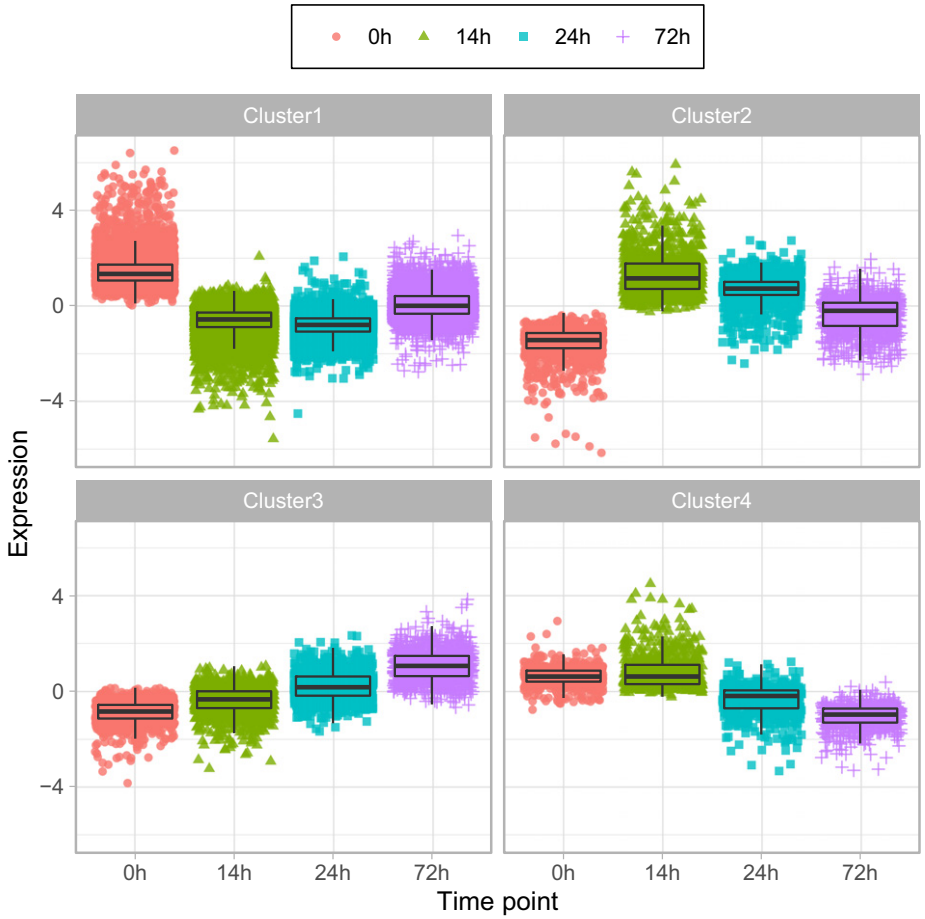


FIG. 2. Boxplot of the DEG clusters in *Phaeocystis antarctica* against sample time points. The boxplots represent the mean expression of each gene in the cluster. These means were calculated for the normalized gene expression values of the replicates at each time point [i.e., log2(mean(fpkm) + 1), scaled to median]. The median is calculated for each time point in each cluster. The shapes indicate the time points. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 4. Sequence features of DEG clusters. The annotated fraction of genes is filtered based on our annotation criteria (see Materials and Methods).

Attribute	Cluster 1 "Immediate-down"	Cluster 2 "Immediate-up"	Cluster 3 "Delayed-up"	Cluster 4 "Delayed-down"
Gene (*n*)	7,774	1,087	1,196	658
Transcript (*n*)	9,134	1,227	1,467	720
Candidate ORF (*n*)	11,612	1,259	1,573	852
GO terms (*n*)	114	24	12	5
eggNOG (*n*)	2,027	276	264	206
Annotated (*n*; %)	2,524; 32%	441; 41%	401; 34%	272; 41%

Carbon metabolism: We identified a large number of differentially expressed genes that were either related to or involved in carbon metabolism (Table H in Appendix S1). Iron addition led to immediate down-regulation of genes related to glucan synthesis (1,3-beta-glucan synthase), glycan degradation (beta-galactosidase), pentose phosphate pathway (PPP), mitochondrial respiration and TCA cycle (malate dehydrogenase), and probably gluconeogenesis (fructose-1,6-bisphosphatase). On the contrary, a delayed down-regulation was observed in glycolysis and glucan catabolism genes (endoglucanase). Contrastingly, iron addition led to immediate up-regulation of one fragment of the Calvin cycle enzyme phosphoribulokinase, whereas another fragment exhibited down- and up-regulation.

A total of 42 lipid metabolism genes were differentially expressed (Table I in Appendix S1). Beta-oxidation-related genes (17) showed immediate (long-chain fatty acid-CoA ligases) or delayed (succinyl-coA ligase) down-regulation following iron addition. Moreover, genes involved in the process of polyunsaturated fatty acid synthesis showed a mixed pattern of up- and down-regulation after iron addition (omega-6 fatty acid desaturase).

Cell cycle, motility and colony formation: In Table J (Appendix S1), 290 genes involved in cell cycle, motility, cytoskeleton structure, and vesicle transport were curated. Interestingly, most of these were down-regulated after iron addition. However, a few genes linked to mucus formation (UDP-O-acylglucosamine N-acyltransferase) showed delayed up-regulation observable only at 72 h. Some down-regulated genes were also involved in actin polymerization, autophagy and membrane remodeling. Mixed patterns of genes involved in cell aggregation were

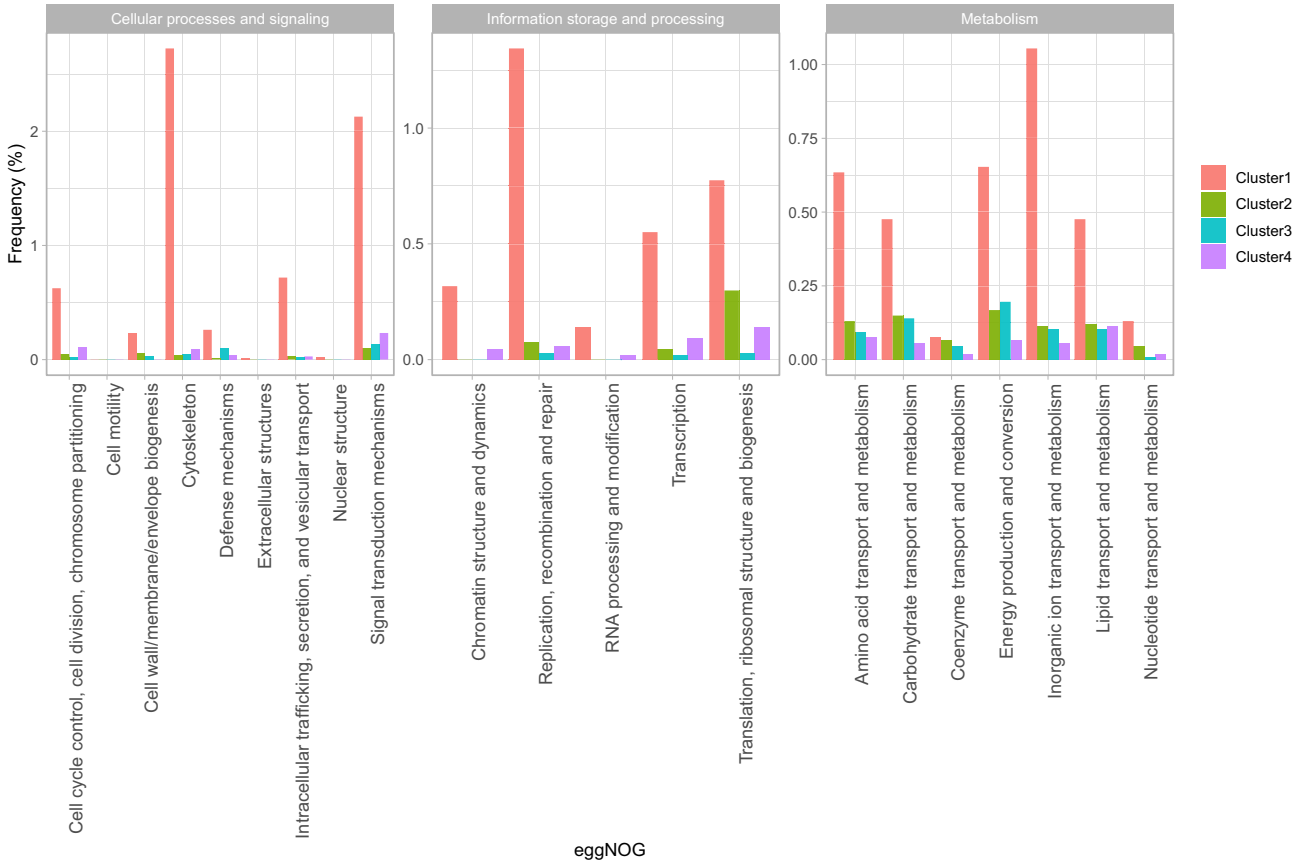


FIG. 3. Representation of differentially expressed eggNOGs in *Phaeocystis antarctica* in four clusters. The frequencies are normalized to the total number of DEGs. [Color figure can be viewed at wileyonlinelibrary.com]

observed (e.g., von Willebrand factor domain, lectin and fibrillin).

DISCUSSION

We conducted a batch culture experiment with iron-limited *Phaeocystis antarctica* in order to assess its acclimation response following iron enrichment at different time points. Using transcriptomics, we monitored the change in expression of key genes and pathways under iron-deplete and iron-replete conditions. Here we aim at providing new insights into the underlying metabolic pathways of adaptation and acclimation to iron enrichment.

Due to the lack of a *Phaeocystis antarctica* genome sequence, we conducted a *de novo* assembly of the transcriptome. We obtained 110,971 genes, a larger number than what was previously reported (56,193 contigs; Koid et al. 2014), and relative to the genome of its sister species *Emiliania huxleyi* (30,569 protein-coding genes; Read et al. 2013). The high number of genes could be attributed to inaccuracy in resolving diploid polymorphisms by Trinity (Grabherr et al. 2011, Haas et al. 2013), a reported phenomenon in diatoms (Armbrust et al. 2004), or alternatively to other factors such as alternative splicing.

We compared the *Phaeocystis antarctica* transcriptome to the published one from the MMETSP project (Koid et al. 2014) in terms of functional and sequence coverage against the BUSCO eukaryote gene set, and the previously published *P. antarctica* plastid and mitochondrial genomes (Smith et al. 2014a), respectively. The transcriptome from our study showed a more complete coverage of BUSCO gene set, and more sequence coverage of both plastid and mitochondrial genomes compared to the MMETSP transcriptome. Thus, our assembly is more comprehensive in regard to functional and sequence coverage. Nevertheless, the low coverage of BUSCO gene set by both our study's transcriptome and that of Koid et al. (2014) could be attributed to the unbalanced representation of the reference plant/algae/fungi sequences in the relatively old version of the eukaryotic gene set. Sequence coverage results as well show a lower degree of overlap between the predicted peptide sequences in the transcriptomic studies than expected, perhaps due to the difference in sequence processing methods and the difference in growth conditions.

The *k*-means-based clustering of the DEGs (Fig. 2) helped identifying patterns in the DEGs qualitatively, however, the increased number of outliers might be attributed to the fusion of clusters especially in cluster 1. The biological significance of the outliers is worth investigating.

Phaeocystis antarctica *acclimates quickly to high iron conditions. Phaeocystis antarctica* is endemic to the largely iron-limited Southern Ocean, which previously responded to added iron with sustained growth in batch culture experiments (Strzepek et al. 2011, Koch et al. 2019). Moreover, its blooms are frequently recorded in the iron-enriched shelf areas (Schoemann et al. 2005, Smith et al. 2014b). We observed 75% of the expressed genes differentially expressed after 14 h, with the majority (79%) being down-regulated. In diatoms, similar acclimation response was reported where the number of iron-limitation–specific genes was larger than that of iron-replete–specific genes (Nunn et al. 2013). Fourteen hours could be considered as a relatively short time with respect to average growth rate under iron limitation (0.3; Strzepek et al. 2011, Koch et al. 2019). However, the observation that the majority of DEGs were down-regulated shortly after iron supply suggests that iron addition perhaps does not evoke a response of its own, but rather restores cellular functions, which were otherwise negatively affected under iron scarcity, for example, by stress alleviation (down-regulation of photoprotection) and restoration of normal cellular functions (up-regulation of photosynthesis-related processes). The down-regulation of a large number of transcript contigs was observed at 24 h as well, along with a peak in RNA translation, perhaps to fuel the up-regulated processes and the increase in cell abundance. At 72 h, a stabilization of gene expression has perhaps been reached and an up-regulation of cell cycle processes has been observed. Below we discuss a number of processes that have been affected under iron enrichment. The change in expression of its key genes is depicted in Figure 4a and schematically represented in Figure 4b.

Iron acquisition and metabolism. Iron plays a ubiquitous role in photosynthetic cells. Its importance stems, on one hand, from the role of iron–sulfur clusters in electron transport (in the chloroplast and the mitochondria; Pilon et al. 2006, Lill 2009). On the other hand, iron is widely used as a cofactor in many other processes such as chlorophyll biosynthesis, assimilation of nitrogen and sulfur, fatty acid metabolism and reactive oxygen species scavenging (Behrenfeld and Milligan 2013, Twining and Baines 2013, Schoffman et al. 2016). Iron requirements of a species can be assessed as the intracellular iron concentration (Strzepek et al. 2011, Twining and Baines 2013). Not surprisingly, SO species—adapted to low iron conditions—have lower iron demands compared to coastal ones (Strzepek et al. 2011), which increase with cell size (Strzepek et al. 2011). As the largest iron supply is required for photosynthetic electron transport, iron is usually concentrated in the plastid (Twining and Baines 2013). To meet their iron requirements, SO species use iron-economic forms of photosynthetic protein complexes (Strzepek et al. 2019), utilize bound iron (Strzepek et al. 2011) and use forms of iron storage (Marchetti et al. 2009).

Down-regulated fragments included genes responsive to iron starvation. Iron-responsive genes have been proposed to have a role in both iron stress sensing and iron acquisition, being mostly surface proteins and sharing iron-dependent regulatory domains (Lommer et al. 2012, Yoshinaga et al. 2014). Specifically, iron-starvation-induced proteins (ISIPs) were universally up-regulated under iron-limiting conditions in subarctic Pacific phytoplankton (Marchetti et al. 2012), suggesting a conserved role in iron uptake (Smith et al. 2016). ISIP2A was found to be activated as an initial response to iron limitation in *Phaeodactylum tricornutum* (Morrissey et al. 2015) and *Thalassiosira oceanica* (Lommer et al. 2012), and was shown to play a role in iron acquisition in the former (McQuaid et al. 2018). Similarly, in our experiment, putative ISIP2A genes showed immediate down-regulation after iron addition, suggesting a role for ISIP2A in iron acquisition also in *P. antarctica* under iron limitation.

The success of low-iron adapted species can be attributed, among other factors, to the molecular alternatives they evolved to utilize bound iron such as siderophore-mediated iron uptake and high-affinity ferric uptake systems (Strzepek et al. 2011, Shaked and Lis 2012, Groussman et al. 2015). *Phaeocystis antarctica* could grow on both organically bound (from ferrichrome and other siderophores; Strzepek et al. 2011) and free iron, with faster rates of iron uptake and ferric iron reduction under low iron conditions (Strzepek et al. 2011). Similar findings were reported in temperate diatoms (e.g., *Phaeodactylum tricornutum,* Morrissey et al. 2015; *Thalassiosira oceanica,* Lommer et al. 2012). These findings illustrate that *P. antarctica* is adapted to utilizing siderophore-bound iron particularly under iron limitation as reported in subantarctic phytoplankton (Maldonado et al. 2005).

Strzepek et al. (2011) demonstrated the existence of a high-affinity ferric reductase-based iron uptake system in *Phaeocystis antarctica.* Such a system consists of a ferric reductase, a multicopper ferroxidase and a permease (Armbrust et al. 2004, Lommer et al. 2012, Morrissey and Bowler 2012). We observed up-regulation of two ferric reduction oxidases and several multicopper oxidases under low iron (Fig. 4a; Table C in Appendix S1), similar to iron-limited *Thalassiosira oceanica* (Lommer et al. 2012). Our results provide further evidence of an iron-regulated ferric uptake system in *P. antarctica.*

To investigate intracellular distribution of iron, we observed immediate and sustained down-regulation of mitoferrin genes after iron addition.

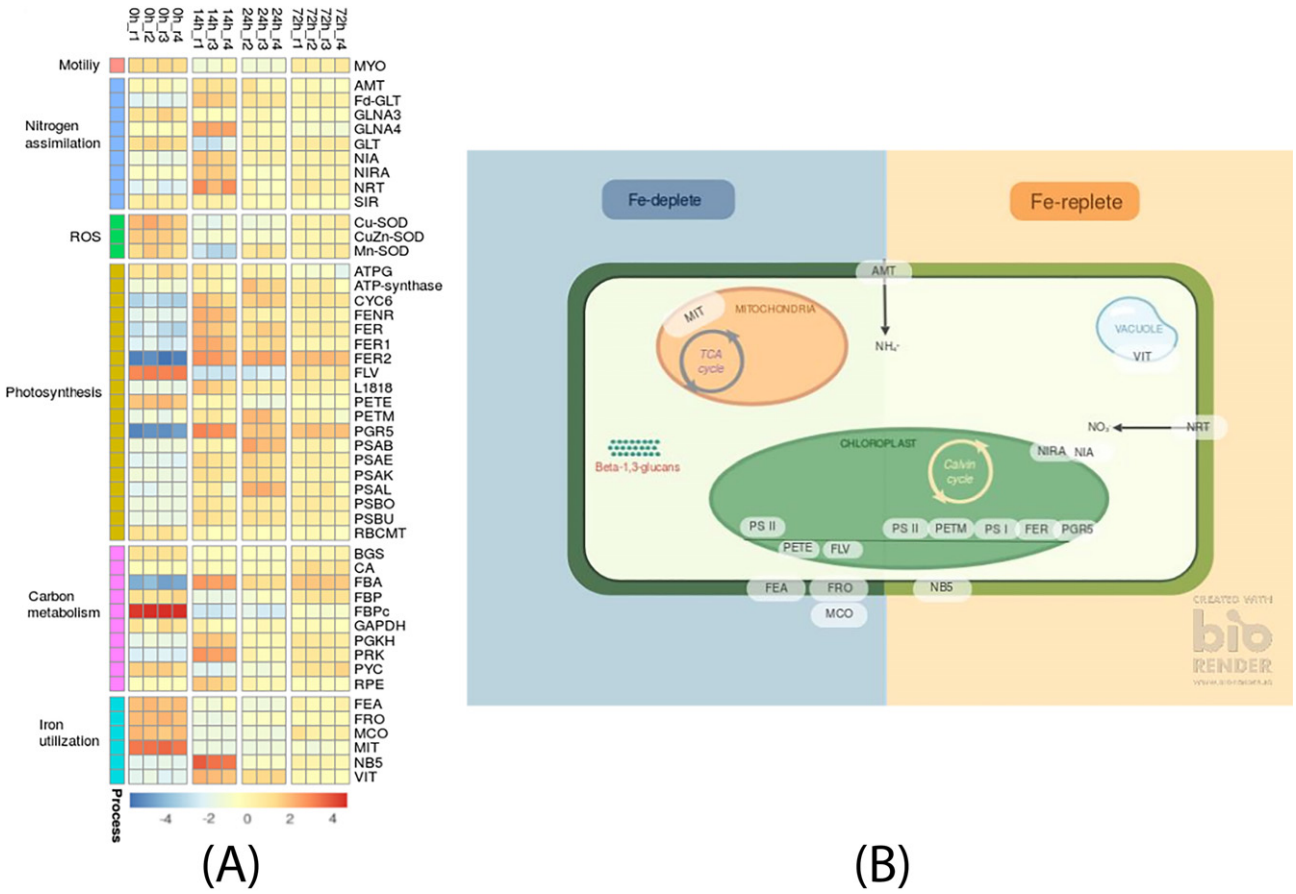


FIG. 4. (A) Heatmap of differentially expressed filtered genes of interest in discussed cellular processes (also in Table M in Appendix S1). Mean normalized expression of genes of the same function is used. (B) Schematic representation of the genes of interest in respect to cellular localization. Gene abbreviations are: AMT, Ammonium transporter 1 members; ATPG, ATP synthase gamma chain; ATP-synthase, ATP synthase subunit a and beta, chloroplastic; BGS, 1,3-beta-glucan synthase components; CA, Carbonic anhydrase 2; Cu-SOD, Cell surface Cu-only superoxide dismutase 5; CuZn-SOD, Cell surface superoxide dismutase [Cu-Zn] 4 and CuZn-SOD chloroplastic; CYC6, Cytochrome c_6; FBA, Fructose-bisphosphate aldolase; FBPc, Fructose-1,6-bisphosphatase, chloroplastic; FBP, Fructose-1,6-bisphosphatase class 1; Fd-GLT, Ferredoxin-dependent glutamate synthase 2; FEA, Low iron-inducible periplasmic protein (ISIP2a domain); FENR, Ferredoxin–NADP reductase, embryo isozyme, chloroplastic; FER1, Ferredoxin-1 and Ferredoxin-1, chloroplastic; FER2, Ferredoxin-2; FER, Ferredoxin; FLV, Flavodoxin; FRO, Ferric reduction oxidase 2 and 6; GAPDH, Glyceraldehyde-3-phosphate dehydrogenase 2; GLNA3, Type-3 glutamine synthetase; GLNA4, Type-3 glutamine synthetase; GLT, Glutamate synthase [NADH], chloroplastic; L1818, Chlorophyll a/b binding protein L1818, chloroplastic; MCO, Multicopper oxidase mco; MIT, Mitoferrin; Mn-SOD, Superoxide dismutase [Mn] and MnSOD mitochondrial; MYO, Myosin and Myosin heavy chain; NB5, NADH-cytochrome b5 reductase 1 and 2; NIA, Nitrate reductase [NADH]; NIRA, Ferredoxin–nitrite reductase, chloroplastic; NRT, High-affinity nitrate transporter 2.4; PETE, Plastocyanin domains; PETM, Cytochrome b6-f complex subunit 7; PGKH, Phosphoglycerate kinase, chloroplastic; PGR5, Protein PROTON GRADIENT REGULATION 5, chloroplastic; PRK, Phosphoribulokinase, chloroplastic; PSAB, Photosystem I P700 chlorophyll a apoprotein A2; PSAE, Photosystem I reaction center subunit IV; PSAK, Photosystem I reaction center subunit PsaK; PSAL, Photosystem I reaction center subunit XI; PSBA, Photosystem II protein D1; PSBO, Oxygen-evolving enhancer protein 1, chloroplastic; PYC, Pyruvate carboxylase; RBCMT, Ribulose-1,5 bisphosphate carboxylase/oxygenase large subunit N-methyltransferase, chloroplastic; RPE, Ribulose-phosphate 3-epimerase; SIR, Sulfite reductase [ferredoxin]; VIT, Vacuolar iron transporter 1 and 1.1.

Mitoferrins are responsible for regulating iron transport in the mitochondria (Shaw et al. 2006), and defects in mitoferrins resulted in impairment in Fe–S cluster assembly and global metabolic changes in land plants (Vigani et al. 2016). Our observation is consistent, on one hand, with the down-regulation of other mitochondrial processes following iron enrichment (e.g., respiration; Table H in Appendix S1). On the other hand, three vacuolar iron transporters and plastidial processes in general (e.g., chloroplastic IscA gene for assembly of Fe–S clusters and photosynthesis) were immediately up-regulated after iron enrichment. These observations suggest that *P. antarctica* shunts iron among compartments depending on iron availability, with iron supplied to the mitochondrium or plastidic processes under low- and high iron conditions, respectively (Fig. 4b).

Cellular metabolism. Photosynthesis is considered a major sink of iron in the cell (Sunda and Huntsman 1995, Strzepek et al. 2011), as iron is integral in the cytochrome b_6f complex and the other components

of the photosynthetic electron transport chain (Strzepek and Harrison 2004). As expected, iron addition led to immediate up-regulation of members of the photosynthetic electron transport chain, pigment production and genes coding for proteins involved in light harvesting; all orchestrated the enhanced photosynthetic efficiency previously observed in our target species (Boyd 2002a, Strzepek et al. 2012, Issak 2014). In particular, flavodoxin and plastocyanin were down-regulated immediately after iron addition. The expression of flavodoxin was proved to be a sign of iron stress in phytoplankton (La Roche et al. 1996) in general and in diatoms in particular (La Roche et al. 1995), with protein expression values elevated at least 25-fold in the diatom *Phaeodactylum tricornutum* under iron limitation (La Roche et al. 1995). This common response was also observed by others in the temperate diatoms *P. tricornutum* (Allen et al. 2008, Zhao et al. 2018) and *Thalassiosira oceanica* (Lommer et al. 2012). Moreover, as an adaptive strategy, a number of SO species, including *Phaeocystis antarctica*, showed a large increase in photosystem II activity under low-iron, low-light, low-temperature conditions that is facilitated by the increased antenna size in those species (Strzepek et al. 2019).

Iron-induced enhancement of photosynthetic capacity logically leads to substantial shifts in carbon metabolism. Given that the production of reducing equivalents through photosynthesis can be hampered by iron limitation (Nunn et al. 2013), glycolysis, TCA cycle and PPP are considered ways to generate reducing equivalents in the form of NAD(P)H when photosynthesis is impaired (Nunn et al. 2013, Rubin et al. 2015). Along these lines, reduced abundance of TCA cycle, glycolysis and PPP-related genes, while higher abundance of Calvin cycle genes were observed in iron-enriched diatoms (Lommer et al. 2012, Nunn et al. 2013). Similarly, iron addition resulted in immediate and sustained up-regulation of genes involved in the Calvin cycle in the tested *Phaeocystis antarctica* strain, and immediate down-regulation of metabolic pathways which recycle or utilize fixed carbon after iron addition.

However, genes related to callose/glucan synthesis were less abundant from the earliest time point on (14 h), those related to glucan catabolism exhibited delayed down-regulation (24 h) after iron addition. Members of the class prymnesiophyceae produce chrysolaminaran (beta-1,3 glucans) as carbon storage products (reviewed in (Alderkamp et al. 2007). Excretion of excess carbon in the form of chrysolaminaran in nutrient-limited *Phaeocystis globosa* was reported (Janse et al. 1996) probably as a vent for excess energy and metabolites under unbalanced growth conditions (Janse et al. 1996, Alderkamp et al. 2007). Accordingly, our results indicate decreased glucan production following iron enrichment. Taken with the aforementioned expression patterns, stored glucan could be remobilized to supply NAD(P)H and/or carbon backbones through carbon recycling/catabolic processes.

However, similar to previous observations in diatoms and haptophytes (Marchetti et al. 2012), fatty acid (FA) degradation genes showed delayed down-regulation after iron addition, FA desaturases exhibited a mixed pattern of down- and up-regulation. FA desaturases contain diiron cluster domains, and different isoforms have different cellular localization (endoplasmic reticulum and plastid), and different electron donors (cytochrome b_5, NADPH and ferredoxin; Sperling and Heinz 2001, Uttaro 2006, Urzica et al. 2013). These results, on one hand, could be explained by a technical limitation in resolving cellular localization and cofactor in our data. On the other hand, they might be an indication of an adaptive strategy, where *Phaeocystis antarctica* activates different FA desaturases according to iron concentration.

Nitrate assimilation is a cellular process that is strongly affected by iron limitation. First, both nitrate and nitrite reductases require iron as a cofactor (Nunn et al. 2013). Second, nitrogen assimilation requires reducing equivalents; the impaired production of which exerts an additional constraint on this process under iron limitation. Interestingly, we observed two different gene expression patterns concerning nitrogen assimilation genes. As expected, nitrate transport and reduction, along with some nitrite reductases (NiRs), glutamate and glutamine synthases were up-regulated after iron addition, while ammonium transporters were down-regulated (Fig. 4a; Table G in Appendix S1). However, other few putative NiRs, glutamate and glutamine synthases and ammonium transporters were down-regulated at the same time, indicating potential adaptation features of tightly regulated acclimation processes. Experiments showed that diatoms utilize urea and ammonia under iron limitation, and when the limitation is alleviated, diatoms switch to nitrate assimilation (Marchetti et al. 2012). In haptophytes, nitrate was the only nitrogen source reported to support both solitary and colonial growth (Wang et al. 2011). Given the importance of the colonial stage to *Phaeocystis antarctica*, it would prefer to maintain a nitrate assimilation activity under low iron. Therefore, it is possible that *P. antarctica*, as with FA desaturases, uses different cofactors for pivotal enzymes under changing iron conditions.

Colony formation in *Phaeocystis antarctica* has been demonstrated to be triggered by increased iron availability (Assmy et al. 2007, Strzepek et al. 2011, Bender et al. 2018). A few molecular markers, suggested to play a role in the formation of the extracellular colonial matrix and in cell aggregation (Bender et al. 2018), were found in colonial *P. antarctica* including von Willebrand domain-containing proteins and adhesin-like proteins. Contrastingly, some fragments exhibiting von Willebrand domains showed elevated expression, whereas most

were down-regulated in our experiment following iron addition. Considering, that colony formation could be observed after 72 h of iron supplementation, the involvement of other fragments in the buildup of mucus and colony skin is implicated as well.

Interestingly, a large number of genes (290) involved in motility and cytoskeleton structures were down-regulated after iron addition (Table J in Appendix S1). The reduced expression of motor proteins such as myosin, dynein and flagellar proteins coincides with an enhanced frequency of colony formation, where the cells lose motility. However, it may also be indicative of lessened phagotrophy and/or membrane trafficking after iron addition, as genes involved in motility, actin polymerization, vesicle transport and possibly adhesion are known to have a role in exocytic/endocytic processes such as secretion or phagocytosis (Rupper and Cardelli 2001). Indeed, ISIP 2A proteins, which concentrate iron on the cell surface, have been shown to be internalized by endocytosis in the diatom *Phaeodactylum tricornutum* (McQuaid et al. 2018). Reduced expression of ISIP 2A genes after iron supplementation could have resulted in reduced frequency of endocytosis in *Phaeocystis antarctica*, along with lessened glucan secretion, leading to the observed gene expression pattern. Alternatively, the reduced expression of genes involved in cytoskeleton structures after iron addition may reflect a reduction in the mixotrophic growth mode. Mixotrophy describes the ability of an organism to use different trophic modes of acquiring macronutrients such as carbon and nitrogen, or trace metals such as iron (Verity et al. 2007, Stoecker et al. 2017, Villanova et al. 2017), allowing for sustained growth even under limiting condition (Stoecker et al. 2017). Evidence of mixotrophy has been found in diatoms (Villanova et al. 2017) and prymnesiophytes (Tillmann 2004, Stoecker et al. 2017). These results call for an investigation of mixotrophic behavior in solitary *P. antarctica* under low iron.

Phaeocystis antarctica *is well adapted to low iron conditions.* There is a number of adaptive strategies of SO phytoplankton species, including *Phaeocystis antarctica*, to combat iron limitation (Strzepek et al. 2011, 2019). In this study, we pinpoint three different possible adaptive strategies.

First, the utilization of iron-free functional alternatives of iron-rich proteins under iron scarcity can be stated as an adaptive approach. This is widely used in temperate and low-iron-adapted organisms (Strzepek and Harrison 2004); including a non-colony-forming *Phaeocystis antarctica* strain (Koch et al. 2019). Examples include the up-regulation of flavodoxin and plastocyanin substituting for ferredoxin and the small heme-containing protein cytochrome c_6, respectively, which we observed in *P. antarctica* immediately after iron addition (Table D in Appendix S1). This response was also observed in haptophytes shortly after iron addition by Marchetti et al. (2012).

Second, an additional strategy to the well-established "iron limitation survival kit" proteins is the activation of iron-dependent and iron-independent isoforms of pivotal metabolic enzymes according to the change in iron conditions (Raven 1988). For example, different isoforms of fructose-bisphosphate aldolase were found in *Thalassiosira oceanica*; operating either with a metal cofactor (class II) or through a Schiff-base catalysis (class I) depending on iron availability (Lommer et al. 2012), allowing for quick acclimation to iron scarce/rich environments. Moreover, flavodoxin was suggested as an electron carrier in plastidic processes such as FA desaturation in iron-limited diatoms (Whitney et al. 2011). We observed differential expression of several genes involved in plastidic pathways such as nitrogen/sulfur assimilation (specifically NiR and GS), and FA desaturation under changing iron conditions (Table G in Appendix S1). Different isoforms of GS and NiR were reported to be expressed in diatoms as well. Metatranscriptomics showed that in diatoms both NADPH-dependent and ferredoxin-dependent GS were up-regulated after iron enrichment (Marchetti et al. 2012), while NADPH-dependent and ferredoxin-dependent NiR were alternately expressed under changing iron conditions (Marchetti et al. 2012). Additionally, no haptophyte nitrate assimilation genes were detected, and it was suggested that haptophytes channel newly acquired iron into photosynthesis rather than nitrate assimilation (Marchetti et al. 2012). Contrastingly, in our experiment, nitrate assimilation genes were differentially expressed under low and high iron in *Phaeocystis antarctica*.

Lastly, we postulate mixotrophy as a possible growth mode of iron-limited solitary *Phaeocystis antarctica* based on enhanced expression of motility and endocytosis-related genes under iron limitation. This feeding mode was also observed in other haptophytes (Tillmann 2004, Stoecker et al. 2017) and was suggested to be active in *Phaeocystis* to overcome limited photosynthetic capacity under prolonged periods of starvation, for instance in winter (Verity et al. 2007), facilitating the uptake of macronutrients and trace elements.

Based on our results, iron addition has led to the up-regulation of photosynthesis genes as well as nitrate assimilation genes simultaneously. It was beyond the scope of this study to resolve which process had the higher priority, a question which may be answered through time points at a finer resolution. However, what can be concluded is that shunting iron toward photosynthesis allows for better photosynthetic efficiency consequently facilitating increased production of reducing equivalents and ATP and therefore a better energy state of the cell. Moreover, as a baseline of the reducing equivalent NADPH is necessary for nitrate assimilation,

Phaeocystis antarctica investing iron into photosynthesis first and nitrogen assimilation second would be a reasonable prioritization of processes, especially if this species is capable of mixotrophic growth.

CONCLUSIONS

Our results suggest that iron-limited *Phaeocystis antarctica* invests in iron acquisition through a Cu-dependent ferric reductase system, and that the majority of iron seems to be directed toward mitochondrial processes. Also, our results demonstrate that *P. antarctica* actively uses a number of adaptive mechanisms to alleviate iron limitation, such as activating iron-economic functional homologs for nitrite reduction and possibly fatty acid biosynthesis. As well, *P. antarctica* uses adaptive strategies such as expressing plastocyanin and flavodoxin under limited iron availability, Finally, the results suggest that *P. antarctica* relies on heterotrophic nutrition through phagocytosis. *Phaeocystis antarctica* responds to iron enrichment by enhancing photosynthetic capacity, a major limiting factor for nitrate assimilation, and consequently perhaps colony formation. A linearity in metabolic changes/shifts in response to added iron was observed in some processes (e.g., iron uptake and transfer) and our observations suggest that the adaptive features enable *Phaeocystis antarctica* to thrive in an environment characterized by chronic iron limitation.

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) in the framework of the priority program "Antarctic Research with comparative investigations in Arctic ice areas" by a grant BE5105/1-1 granted to SB, and by a DFG grant (131153660) to SG. ST was funded by the Helmholtz association (Young Investigator Group EcoTrace, VH-NG-901). MRR was funded by Al Alfi Foundation for Human and Social Development. We thank Dieter Wolf-Gladrow, Christine Klaas, Katja Metfies, Klaus Valentin, and Susana C. Vazquez for their valuable input to the study and manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

SB and ST were involved in study and experimental design. SB and MRR were involved in experiment and wet laboratory work. SF, SB, and MRR were involved in conceptualization. MRR, LH, VB, and SB were involved in data curation. MRR, LH, SF, and AM were involved in formal analysis and software. SB, ST, SG, and MRR were involved in funding and resources. MRR, SF, AM, and SB were involved in investigation and original draft preparation. MRR, AM, and SF were involved in visualization. MRR, AM, SB, SF, ST, SG, VB, and LH were involved in review and editing.

Alderkamp, A. C., Buma, A. G. J. & van Rijssel, M. 2007. The carbohydrates of *Phaeocystis* and their degradation in the microbial food web. *Biogeochemistry* 83:99–118.

Alderkamp, A. C., Kulk, G., Buma, A. G. J., Visser, R. J. W., Van Dijken, G. L., Mills, M. M. & Arrigo, K. R. 2012. The effect of iron limitation on the photophysiology of *Phaeocystis antarctica* (Prymnesiophyceae) and *Fragilariopsis cylindrus* (Bacillariophyceae) under dynamic irradiance. *J. Phycol.* 48:45–59.

Allen, A. E., La Roche, J., Maheswari, U., Lommer, M., Schauer, N., Lopez, P. J., Finazzi, G., Fernie, A. R. & Bowler, C. 2008. Whole-cell response of the pennate diatom *Phaeodactylum tricornutum* to iron starvation. *Proc. Natl. Acad. Sci. USA* 105:10438–43.

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S. et al. 2004. The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. *Science* 306:79–86.

Assmy, P., Henjes, J., Klaas, C. & Smetacek, V. 2007. Mechanisms determining species dominance in a phytoplankton bloom induced by the iron fertilization experiment EisenEx in the Southern Ocean. *Deep Sea Res. Part I Oceanogr. Res. Pap.* 54:340–62.

de Baar, H. J. W., Boyd, P. W., Coale, K. H., Landry, M. R., Tsuda, A., Assmy, P., Bakker, D. C. E. et al. 2005. Synthesis of iron fertilization experiments: From the iron age in the age of enlightenment. *J. Geophys. Res.* 110:C09S16.

Beardall, J., Allen, D., Bragg, J., Finkel, Z. V., Flynn, K. J., Quigg, A., Rees, T. A. V., Richardson, A. & Raven, J. A. 2009. Allometry and stoichiometry of unicellular, colonial and multicellular phytoplankton. *New Phytol.* 181:295–309.

Behrenfeld, M. J. & Milligan, A. J. 2013. Photophysiological expressions of iron stress in phytoplankton. *Ann. Rev. Mar. Sci.* 5:217–46.

Bender, S. J., Moran, D. M., McIlvin, M. R., Zheng, H., McCrow, J. P., Badger, J., DiTullio, G. R., Allen, A. E. & Saito, M. A. 2018. Colony formation in *Phaeocystis antarctica*: Connecting molecular mechanisms with iron biogeochemistry. *Biogeosciences* 15:4923–42.

Beszteri, S., Yang, I., Jaeckisch, N., Tillmann, U., Frickenhaus, S., Glöckner, G., Cembella, A. & John, U. 2012. Transcriptomic response of the toxic prymnesiophyte *Prymnesium parvum* (N. Carter) to phosphorus and nitrogen starvation. *Harmful Algae* 18:1–15.

Bolger, A. M., Lohse, M. & Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–20.

Borowitzka, M. A. 2018. The 'stress' concept in microalgal biology —homeostasis, acclimation and adaptation. *J. Appl. Phycol.* 30:2815–25.

Boyd, P. W. 2002a. Environmental factors controlling phytoplankton processes in the Southern Ocean. *J. Phycol.* 38:844–61.

Boyd, P. W. 2002b. The role of iron in the biogeochemistry of the Southern Ocean and equatorial Pacific: a comparison of in situ iron enrichments. *Deep Sea Res. Part II Top. Stud. Oceanogr.* 49:1803–21.

Boyd, P. W., Arrigo, K. R., Strzepek, R. & van Dijken, G. L. 2012. Mapping phytoplankton iron utilization: Insights into Southern Ocean supply mechanisms. *J. Geophys. Res.* 117: C06009.

Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. 2014. NbClust: an R package for determining the relevant number of clusters in a data set. *J. Stat. Softw.* 61:1–36.

Dugdale, R. C. & Wilkerson, F. P. 1991. Low specific nitrate uptake rate: A common feature of high-nutrient, low-chlorophyll marine ecosystems. *Limnol. Oceanogr.* 36:1678–88.

Gaebler-Schwarz, S., Davidson, A., Assmy, P., Chen, J., Henjes, J., Nöthig, E. M., Lunau, M. & Medlin, L. K. 2010. A new cell stage in the haploid-diploid life cycle of the colony-forming haptophyte *Phaeocystis antarctica* and its ecological implications. *J. Phycol.* 46:1006–16.

Gall, M., Boyd, P., Hall, J., Safi, K. & Chang, H. 2001. Phytoplankton processes. Part 1: Community structure during the Southern Ocean Iron RElease Experiment (SOIREE). *Deep Sea Res. Part II Top. Stud. Oceanogr.* 48:2551–70.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X. et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29:644–52.

Groussman, R. D., Parker, M. S. & Armbrust, E. V. 2015. Diversity and evolutionary history of iron metabolism genes in diatoms. *PLoS ONE* 10:e0129081.

Guillard, R. R. L. & Ryther, J. H. 1962. Studies of marine planktonic diatoms. I. *Cyclotella nana* Hustedt, and *Detonula confervacea* (Cleve) Gran. *Can. J. Microbiol.* 8:229–39.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B. et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8:1494–512.

Hamm, C. 2000. Architecture, ecology and biogeochemistry of *Phaeocystis* colonies. *J. Sea Res.* 43:307–15.

Harel, A., Bromberg, Y., Falkowski, P. G. & Bhattacharya, D. 2014. Evolutionary history of redox metal-binding domains across the tree of life. *Proc. Natl. Acad. Sci. USA* 111:7042–7.

Hutchins, D. A. & Boyd, P. W. 2016. Marine phytoplankton and the changing ocean iron cycle. *Nat. Clim. Chang.* 6:1072–9.

Issak, M. R. R. 2014. Transcriptomics of iron limitation in *Phaeocystis antarctica*. The American University in Cairo, 109 pp.

Janse, I., Rijssel, M., Hall, P., Gerwig, G. J., Gottschal, J. C. & Prins, R. A. 1996. The storage glucan of *Phaeocystis globosa* (Prymnesiophyceae) cells. *J. Phycol.* 32:382–7.

Koch, F., Beszteri, S., Harms, L. & Trimborn, S. 2019. The impacts of iron limitation and ocean acidification on the cellular stoichiometry, photophysiology, and transcriptome of *Phaeocystis antarctica*. *Limnol. Oceanogr.* 64:357–75.

Koid, A. E., Liu, Z., Terrado, R., Jones, A. C., Caron, D. A. & Heidelberg, K. B. 2014. Comparative transcriptome analysis of four prymnesiophyte algae. *PLoS ONE* 9:e97801.

La Roche, J., Boyd, P. W., McKay, R. M. L. & Geider, R. J. 1996. Flavodoxin as an in situ marker for iron stress in phytoplankton. *Nature* 382:802–5.

La Roche, J., Murray, H., Orellana, M. & Newton, J. 1995. Flavodoxin expression as an indicator of iron limitation in marine diatoms. *J. Phycol.* 31:520–30.

Li, B. & Dewey, C. N. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323.

Li, L., Stoeckert, C. J. & Roos, D. S. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13:2178–89.

Lill, R. 2009. Function and biogenesis of iron – sulphur proteins. *Nature* 460:831–8.

Lommer, M., Specht, M., Roy, A. S., Kraemer, L., Andreson, R., Gutowska, M. A., Wolf, J. et al. 2012. Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biol.* 13:R66.

Love, M. I., Huber, W. & Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550.

Maldonado, M. T., Strzepek, R. F., Sander, S. & Boyd, P. W. 2005. Acquisition of iron bound to strong organic complexes, with different Fe binding groups and photochemical reactivities, by plankton communities in Fe-limited subantarctic waters. *Global Biogeochem. Cy.* 19:GB4S23.

Marchetti, A., Parker, M. S., Moccia, L. P., Lin, E. O., Arrieta, A. L., Ribalet, F., Murphy, M. E. P., Maldonado, M. T. & Armbrust, E. V. 2009. Ferritin is used for iron storage in bloom-forming marine pennate diatoms. *Nature* 457:467–70.

Marchetti, A., Schruth, D. M., Durkin, C. A., Parker, M. S., Kodner, R. B., Berthiaume, C. T., Morales, R., Allen, A. E. & Armbrust, E. V. 2012. Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proc. Natl. Acad. Sci. USA* 109:E317–25.

Martin, J. H., Gordon, R. M. & Fitzwater, S. E. 1990. Iron in Antarctic waters. *Nature* 345:156–8.

McQuaid, J. B., Kustka, A. B., Oborník, M., Horák, A., McCrow, J. P., Karas, B. J., Zheng, H., Kindeberg, T., Andersson, A. J., Barbeau, K. A. & Allen, A. E. 2018. Carbonate-sensitive phytotransferrin controls high-affinity iron uptake in diatoms. *Nature* 555:534–7.

Morrissey, J. & Bowler, C. 2012. Iron utilization in marine cyanobacteria and eukaryotic algae. *Front. Microbiol.* 3:43.

Morrissey, J., Sutak, R., Paz-Yepes, J., Tanaka, A., Moustafa, A., Veluchamy, A., Thomas, Y. et al. 2015. A novel protein, ubiquitous in marine phytoplankton, concentrates iron at the cell surface and facilitates uptake. *Curr. Biol.* 25:364–71.

Na, D., Son, H. & Gsponer, J. 2014. Categorizer: a tool to categorize genes into user-defined biological groups based on semantic similarity. *BMC Genom.* 15:1091.

Neumann, R., Kumar, S., Haverkamp, T. H. & Shalchian-Tabrizi, K. 2014. BLASTGrabber: a bioinformatic tool for visualization, analysis and sequence selection of massive BLAST data. *BMC Bioinformatics* 15:128.

Nunn, B. L., Faux, J. F., Hippmann, A. A., Maldonado, M. T., Harvey, H. R., Goodlett, D. R., Boyd, P. W. & Strzepek, R. F. 2013. Diatom proteomics reveals unique acclimation strategies to mitigate Fe limitation. *PLoS ONE* 8:e75653.

Pilon, M., Abdel-Ghany, S. E., Hoewyk, D., Ye, H. & Pilon-Smits, E. A. H. 2006. Biogenesis of iron-sulfur cluster proteins in plastids. *In* Setlow, J. K. [Ed.] *Genetic Engineering: Principles and Methods*. Springer, Boston, MA, pp. 101–17.

Raven, J. 1988. The iron and molybdenum use efficiencies of plant growth with different energy, carbon and nitrogen sources. *New Phytol.* 109:279–87.

Raven, J. A. 2013. Iron acquisition and allocation in stramenopile algae. *J. Exp. Bot.* 64:2119–27.

Read, B. A., Kegel, J., Klute, M. J., Kuo, A., Lefebvre, S. C., Maumus, F., Mayer, C. et al. 2013. Pan genome of the phytoplankton *Emiliania* underpins its global distribution. *Nature* 499:209–13.

Rubin, B. E., Wetmore, K. M., Price, M. N., Diamond, S., Shultzaberger, R. K., Lowe, L. C., Curtin, G., Arkin, A. P., Deutschbauer, A. & Golden, S. S. 2015. The essential gene set of a photosynthetic organism. *Proc. Natl. Acad. Sci. USA* 112: E6634–43.

Rupper, A. & Cardelli, J. 2001. Regulation of phagocytosis and endo-phagosomal trafficking pathways in *Dictyostelium discoideum*. *Biochim. Biophys. Acta – Gen. Subj.* 1525:205–16.

Schoemann, V., Becquevort, S., Stefels, J., Rousseau, V. & Lancelot, C. 2005. *Phaeocystis* blooms in the global ocean and their controlling mechanisms: a review. *J. Sea Res.* 53:43–66.

Schoffman, H., Lis, H., Shaked, Y. & Keren, N. 2016. Iron-nutrient interactions within phytoplankton. *Front. Plant Sci.* 7:1223.

Shaked, Y. & Lis, H. 2012. Disassembling iron availability to phytoplankton. *Front. Microbiol.* 3:123.

Shaw, G. C., Cope, J. J., Li, L., Corson, K., Hersey, C., Ackermann, G. E., Gwynn, B. et al. 2006. Mitoferrin is essential for erythroid iron assimilation. *Nature* 440:96–100.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–2.

Smetacek, V., De Baar, H. J. W., Bathmann, U. V., Lochte, K. & Rutgers Van Der Loeff, M. M. 1997. Ecology and biogeochemistry of the Antarctic circumpolar current during austral spring: a summary of southern ocean JGOFS cruise ANT X/6 of R.V. Polarstern. *Deep Sea Res. Part II Top. Stud. Oceanogr.* 44:1–21.

Smith, W. O., Ainley, D. G., Arrigo, K. R. & Dinniman, M. S. 2014b. The oceanography and ecology of the Ross Sea. *Ann. Rev. Mar. Sci.* 6:469–87.

Smith, D. R., Arrigo, K. R., Alderkamp, A. C. & Allen, A. E. 2014a. Massive difference in synonymous substitution rates

among mitochondrial, plastid, and nuclear genes of *Phaeocystis* algae. *Mol. Phylogenet. Evol.* 71:36–40.
Smith, S. R., Gillard, J. T. F., Kustka, A. B., McCrow, J. P., Badger, J. H., Zheng, H., New, A. M., Dupont, C. L., Obata, T., Fernie, A. R. & Allen, A. E. 2016. Transcriptional orchestration of the global cellular response of a model pennate diatom to diel light cycling under iron limitation. *PLoS Genet.* 12:e1006490.
Sperling, P. & Heinz, E. 2001. Desaturases fused to their electron donor. *Eur. J. Lipid Sci. Technol.* 103:158–80.
Stoecker, D. K., Hansen, P. J., Caron, D. A. & Mitra, A. 2017. Mixotrophy in the marine plankton. *Ann. Rev. Mar. Sci.* 9:311–35.
Strzepek, R. F., Boyd, P. W. & Sunda, W. G. 2019. Photosynthetic adaptation to low iron, light, and temperature in Southern Ocean phytoplankton. *Proc. Natl. Acad. Sci. USA* 116:4388–93.
Strzepek, R. F. & Harrison, P. J. 2004. Photosynthetic architecture differs in coastal and oceanic diatoms. *Nature* 431: 689–92.
Strzepek, R. F., Hunter, K. A., Frew, R. D., Harrison, P. J. & Boyd, P. W. 2012. Iron-light interactions differ in Southern Ocean phytoplankton. *Limnol. Oceanogr.* 57:1182–200.
Strzepek, R. F., Maldonado, M. T., Hunter, K. A., Frew, R. D. & Boyd, P. W. 2011. Adaptive strategies by Southern Ocean phytoplankton to lessen iron limitation: uptake of organically complexed iron and reduced cellular iron requirements. *Limnol. Oceanogr.* 56:1983–2002.
Sunda, W. G. & Huntsman, S. A. 1995. Iron uptake and growth limitation in oceanic and coastal phytoplankton. *Mar. Chem.* 50:189–206.
Suzuki, R. & Shimodaira, H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22:1540–2.
Tillmann, U. 2004. Interactions between planktonic microalgae and protozoan grazers. *J. Eukaryot. Microbiol.* 51:156–68.
Trimborn, S., Brenneis, T., Hoppe, C., Laglera, L., Norman, L., Santos-Echeandía, J., Völkner, C., Wolf-Gladrow, D. & Hassler, C. 2017. Iron sources alter the response of Southern Ocean phytoplankton to ocean acidification. *Mar. Ecol. Prog. Ser.* 578:35–50.
Twining, B. S. & Baines, S. B. 2013. The trace metal composition of marine phytoplankton. *Ann. Rev. Mar. Sci.* 5:191–215.
Urzica, E. I., Vieler, A., Hong-Hermesdorf, A., Page, M. D., Casero, D., Gallaher, S. D., Kropat, J., Pellegrini, M., Benning, C. & Merchant, S. S. 2013. Remodeling of membrane lipids in iron-starved *Chlamydomonas*. *J. Biol. Chem.* 288:30246–58.
Uttaro, A. 2006. Biosynthesis of polyunsaturated fatty acids in lower eukaryotes. *IUBMB Life* 58:563–71.
Verity, P. G., Brussaard, C. P., Nejstgaard, J. C., Leeuwe, M. A., Lancelot, C. & Medlin, L. K. 2007. Current understanding of *Phaeocystis* ecology and biogeochemistry, and perspectives for future research. *Biogeochemistry* 83:311–30.
Vigani, G., Bashir, K., Ishimaru, Y., Lehmann, M., Casiraghi, F. M., Nakanishi, H., Seki, M., Geigenberger, P., Zocchi, G. & Nishizawa, N. K. 2016. Knocking down mitochondrial iron transporter (MIT) reprograms primary and secondary metabolism in rice plants. *J. Exp. Bot.* 67:1357–68.
Villanova, V., Fortunato, A. E., Singh, D., Bo, D. D., Conte, M., Obata, T., Jouhet, J. et al. 2017. Investigating mixotrophic metabolism in the model diatom *Phaeodactylum tricornutum*. *Philos. T. Roy. Soc. B.* 372:20160404.
Wang, X., Wang, Y. & Smith, W. O. 2011. The role of nitrogen on the growth and colony development of Phaeocystis globosa (Prymnesiophyceae). *Eur. J. Phycol.* 46:305–14.
Whitney, L. P., Lins, J. J., Hughes, M. P., Wells, M. L., Chappell, P. D. & Jenkins, B. D. 2011. Characterization of putative iron responsive genes as species-specific indicators of iron stress in thalassiosiroid diatoms. *Front. Microbiol.* 2:234.
Yoshinaga, R., Niwa-Kubota, M., Matsui, H. & Matsuda, Y. 2014. Characterization of iron-responsive promoters in the marine diatom *Phaeodactylum tricornutum*. *Mar. Genomics* 16:55–62.
Zhao, P., Gu, W., Huang, A., Wu, S., Liu, C., Huan, L., Gao, S., Xie, X. & Wang, G. 2018. Effect of iron on the growth of *Phaeodactylum tricornutum* via photosynthesis. *J. Phycol.* 54:34–43.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web site:

Table S1. Constituents of f/2 phytoplankton growth medium modified from Guillard and Ryther (1962) and their final concentrations.

Table S2. Physiological measurements, and iron and light status of treatment cultures before and after iron supplementation (mean ± SD, n = 4).

Table S3. Gene families identified using eggNOG (≥ 100 genes per family).

Figure S1. Frequency distribution of *Phaeocystis antarctica* transcripts (i.e., isoforms) counts per genes (log transformed).

Figure S2. Correlation clustering of experiment samples (i.e., replicates) performed on gene counts**.** The cluster was done by Pvclust (nboot = 1000; AU: Approximately Unbiased p-value; BP: Bootstrap Probability value). The bootstrap clustering has been confirmed with correlation clustering based on Spearman correlation, hierarchical clustering of normalized gene expression based on sample correlation and principal component analysis (data not shown). Principal component analysis showed replicate 1 at 24 h as outlier and we eliminated it from downstream differential expression analysis.

Figure S3. *K*-mean clustering of DEGs in *Phaeocystis antarctica* against time. The gray lines represent the mean expression of each gene. These means were calculated for the normalized gene expression values of the replicates at each timepoint (i.e., log2[mean(fpkm) + 1], scaled to median). The colored dots represent the mean expression profile of the DEGs within a cluster.

Appendix S1. Analysis of DEGs. Results of GO enrichment analysis of DEGs in table A, eggNOG categories in table B, and biological process-specific DEGs in tables C–M.

B.2 Detection of drug risks after approval: Methods development for the use of routine statutory health insurance data

Contribution to the manuscript: R. Foraita planned the concept and wrote the main draft of the publication. I and R. Foraita conducted the literature research, prepared visualizations and wrote the initial draft of the section concerning using functional targets-based analysis and construction of patient risk profiles in identifying risk factors for adverse drug events in routine statutory health insurance data.

Bundesgesundheitsbl 2018 · 61:1075–1081
https://doi.org/10.1007/s00103-018-2786-z
Online publiziert: 19. Juli 2018

Ronja Foraita[1] · Louis Dijkstra[1] · Felix Falkenberg[2] · Marco Garling[2] · Roland Linder[2] · René Pflock[3] · Mariam R. Rizkallah[1] · Markus Schwaninger[3] · Marvin N. Wright[1] · Iris Pigeot[1]

[1] Leibniz-Institut für Präventionsforschung und Epidemiologie – BIPS, Bremen, Deutschland

[2] Wissenschaftliches Institut der Techniker Krankenkasse für Nutzen und Effizienz im Gesundheitswesen (WINEG TK), Hamburg, Deutschland

[3] Institut für Experimentelle und Klinische Pharmakologie und Toxikologie, Universität zu Lübeck, Lübeck, Deutschland

Aufdeckung von Arzneimittelrisiken nach der Zulassung

Methodenentwicklung zur Nutzung von Routinedaten der gesetzlichen Krankenversicherungen

Einleitung

Basierend auf den Zahlen einer Metaanalyse aus den USA wurde geschätzt, dass unerwünschte Arzneimittelwirkungen (UAW) zu den 4–6 häufigsten Todesursachen in den USA zählen [1]. Einem jüngeren Bericht der Europäischen Kommission zufolge sind europaweit jährlich 100.800–197.000 Todesfälle und ca. 3–10 % der Krankenhauseinweisungen auf UAW zurückzuführen [2]. Bei älteren Patienten wird der Anteil der Krankenhauseinweisungen aufgrund von UAW sowohl auf europäischer Ebene als auch weltweit auf 5–10 % geschätzt [3, 4]. Ein ähnlicher Anteil wird auch für Deutschland berichtet [5]. Immer wieder kommt es zu Marktrücknahmen auch häufig verwendeter Arzneimittel aus Sicherheitsgründen, da schwerwiegende UAW in den klinischen Studien vor der Zulassung nicht erkannt wurden. Beispielsweise geht man davon aus, dass in dem Zeitraum, in dem Vioxx (Merck, New Jersey, USA; Wirkstoff Rofecoxib) verschrieben wurde, allein in Deutschland mehrere Tausend Personen UAW (u. a. Myokardinfarkte) erfahren haben [6, 7]. Aufgabe der Pharmakovigilanz ist es, durch die systematische Überwachung von Arzneimitteln nach der Zulassung solche zum Zeitpunkt der Zulassung noch unbekannten Risiken aufzudecken.

Die Pharmakovigilanz in Europa, wie auch in vielen anderen Ländern, beruht primär auf spontanen Verdachtsmeldungen von einer möglichen UAW. Zur Sammlung dieser Fallberichte wurde gemäß der 2012 in Kraft getretenen Gesetzgebung der Europäischen Union ein bei der europäischen Arzneimittelagentur (EMA) angesiedeltes, zentrales Spontanmelderegister von möglichen UAW, die sog. EudraVigilance-Datenbank, geschaffen. An dieses Register müssen alle Verdachtsfälle von Arzneimittelnebenwirkungen durch die pharmazeutische Industrie gemeldet werden. Ärztinnen und Ärzte in Deutschland sind per Berufsordnung verpflichtet, die ihnen aus ihrer ärztlichen Behandlungstätigkeit bekannt werdenden UAW der Arzneimittelkommission der deutschen Ärzteschaft mitzuteilen [8]. Dieses Fachgremium leitet die Meldungen an das Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) weiter, das als zuständige Bundesoberbehörde die Daten in EudraVigilance einpflegt [9, 10]. Im November 2017 wurde dementsprechend das in Deutschland am BfArM angesiedelte Register geschlossen und in die EudraVigilance-Datenbank überführt. Die in dieser Datenbank kumulierten Informationen zu Expositionen (Arzneimitteln) und Ereignissen (vermuteten UAW) werden anhand speziell entwickelter Algorithmen analysiert, um potenzielle Sicherheitsrisiken („Signale“) zu entdecken. Eine vereinfachte Darstellung des Prozesses zur Signalerkennung findet sich in **Abb. 1**, in der auch einige gängige Verfahren zur Signalerkennung aufgeführt sind, auf die in den nächsten Abschnitten zum Teil noch eingegangen wird.

Allerdings unterliegen Spontanmeldedaten zur Identifizierung potenzieller Sicherheitsrisiken einigen in der Literatur ausführlich dokumentierten Limitationen [12, 13]. So werden nur ca. 5–10 % der Arzneimittelwirkungen tatsächlich gemeldet, wodurch sich ein erhebliches „Underreporting“ ergibt. Dies betraf auch den Verdacht auf mögliche UAW bedingt durch Vioxx, der erst durch die Nutzung von Routinedaten für Pharmakovigilanzzwecke aufkam [6]. Umgekehrt kann es aber auch zu einem „Overreporting“ kommen, wenn ein Ereignis von verschiedenen Stellen mehrfach gemeldet wird. Ein grundlegendes Problem ergibt sich zudem dadurch, dass in Spontanmelderegistern die Anzahl der Exponierten unbekannt ist, sodass die relative Häufigkeit von Ereignissen nicht ermittelt werden kann [14, 15].

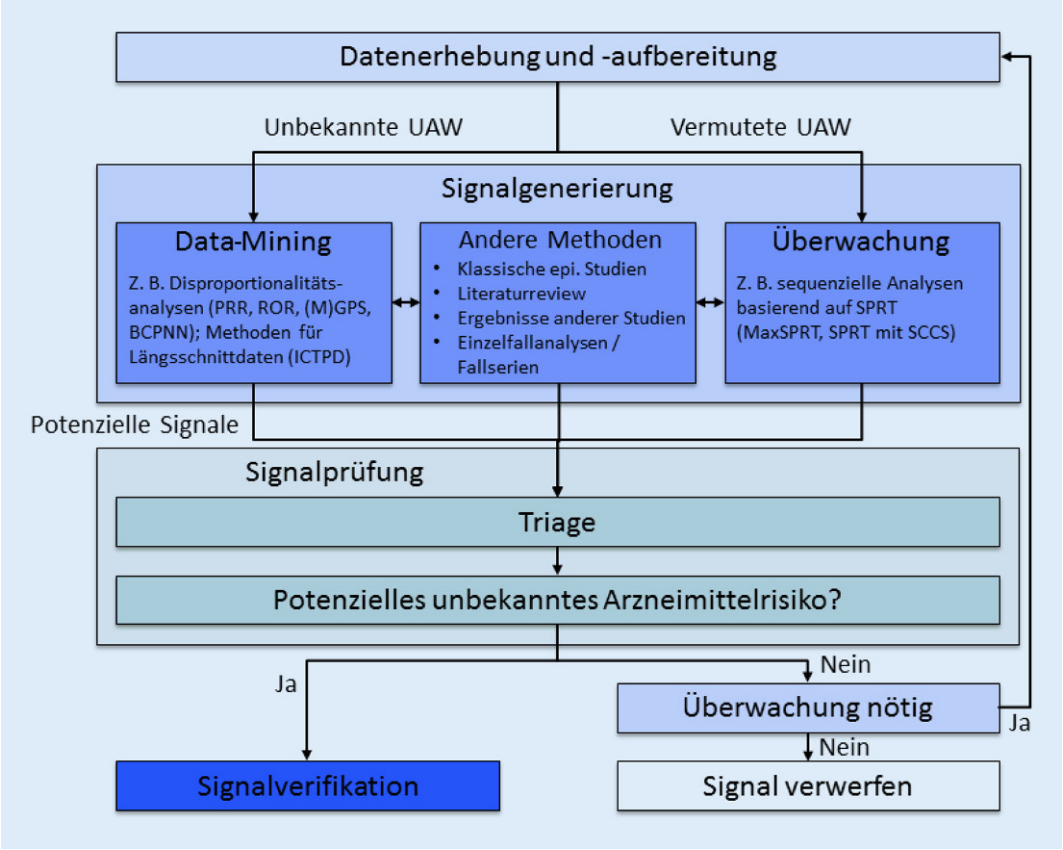


Abb. 1 ◄ Schematische Darstellung des Prozesses zur Signalerkennung. *UAW* Unerwünschte Arzneimittelwirkung, *PRR* Proportional Reporting Ratio, *ROR* Reporting Odds Ratio, *(M)GPS* (Multiitem) Gamma-Poisson Shrinker, *BCPNN* Bayesian Confidence Propagation Neural Network, *ICTPD* Information Component Temporal Pattern Discovery, *SPRT* Sequential Probability Ratio Test, *SCCS* Self-controlled Case Series. (Übersetzte Abbildung aus [11], © M. Suling, I. Pigeot. Die Abbildung ist lizenziert unter der Creative Commons Attribution License 3.0 [https://creativecommons.org/licenses/by/3.0/]).

Vor diesem Hintergrund stellt die Etablierung eines Systems zur Untersuchung der Arzneimittelsicherheit (als Ergänzung zu Spontanmelderegistern), das auf der Nutzung von Versichertendaten basiert, ein wertvolles Instrument dar, das einen bedeutenden Beitrag zur Patientensicherheit in der Versorgung leisten kann [16, 17]. Die Routinedaten bieten Informationen zu den abgegebenen Arzneimitteln und zum Auftreten von Diagnosen mit einer kalendarischen Zeitangabe [18]. Insbesondere Krankenhauseinweisungen nach Arzneimittelverschreibung können als Informationen über mögliche schwere UAW dienen.

In diesem methodischen Artikel werden Verfahren zur Signalerkennung in Abrechnungsdaten der gesetzlichen Krankenversicherungen (GKV) vorgestellt, wobei schwerpunktmäßig neue Konzepte diskutiert werden. Diese sollen dazu beitragen, drei Kernprobleme von Arzneimittelsicherheitsstudien zu lösen: (1) Verminderung der Anzahl falsch-positiver Signale, (2) Identifikation seltener Risiken und (3) Identifikation von Bevölkerungsgruppen mit erhöhtem Risiko. Als zentrale Datenbank wird die deutsche pharmakoepidemiologische Forschungsdatenbank (GePaRD) herangezogen, die zurzeit bundesweite Abrechnungsdaten von mehr als 24 Mio. Versicherten von vier GKVen der Jahre 2004 bis 2015 umfasst (u. a. [18]). In der abschließenden Diskussion wird zusammenfassend aufgezeigt, wie die verschiedenen Methoden der Signalerkennung zum Nutzen potenzieller Betroffener eingesetzt werden können.

Methoden der Signalerkennung

Die statistischen Methoden der Pharmakovigilanz wurden hauptsächlich für die Auswertung von Spontanmeldedaten entwickelt. Um die Vorteile von Routinedaten zu nutzen, werden statistische Methoden weiterentwickelt, die sich bei der Auswertung von sehr großen und strukturierten Datenmengen (wie z. B. bei genetischen Auswertungen) bewährt haben.

Methoden für Spontanmeldedaten und ihr Potenzial für Abrechnungsdaten der Krankenkassen

Die Hauptaufgabe der Pharmakovigilanz ist die Detektion von bisher unbekannten Assoziationen zwischen Arzneimitteln und UAW. Die folgende Darstellung der Methoden der Pharmakovigilanz folgt dem Prozess der Datenaufarbeitung bis hin zu den berichteten Signalen (**Abb. 1**): Zu Beginn werden die in dem Register vorliegenden Meldungen zu Arzneimitteln und UAW für eine weitergehende Verarbeitung (Schritt: Signalgenerierung) aufbereitet. Bei der anschließenden Signalprüfung werden die vorgefundenen UAW-Meldungen, die in Zusammenhang mit einem Medikament stehen, gelistet, auf medizinische Plausibilität überprüft und nach Schweregrad priorisiert (Triage). Den Abschluss des Prozesses stellt die Entscheidung dar, welche Schritte der Signalprüfung folgen sollen: Entweder wird das Sicherheitsrisiko des Signals als so hoch eingestuft, dass umgehend eine konfirmatorische Studie

Zusammenfassung · Abstract

Bundesgesundheitsbl 2018 · 61:1075–1081 https://doi.org/10.1007/s00103-018-2786-z

R. Foraita · L. Dijkstra · F. Falkenberg · M. Garling · R. Linder · R. Pflock · M. R. Rizkallah · M. Schwaninger · M. N. Wright · I. Pigeot

Aufdeckung von Arzneimittelrisiken nach der Zulassung. Methodenentwicklung zur Nutzung von Routinedaten der gesetzlichen Krankenversicherungen

Zusammenfassung

Unerwünschte Arzneimittelwirkungen zählen zu den häufigen Todesursachen. Aufgabe der Pharmakovigilanz ist es, Arzneimittel nach der Zulassung zu überwachen, um so mögliche Risiken aufzudecken. Zu diesem Zweck werden typischerweise Spontanmelderegister genutzt, an die u. a. Ärzte und pharmazeutische Industrie Berichte über unerwünschte Arzneimittelwirkungen (UAW) melden. Diese Register sind jedoch nur begrenzt geeignet, um potenzielle Sicherheitsrisiken zu identifizieren. Eine andere, möglicherweise informativere Datenquelle sind Abrechnungsdaten der gesetzlichen Krankenversicherungen (GKV), die nicht nur den Gesundheitszustand eines Patienten im Längsschnitt erfassen, sondern auch Informationen zu Begleitmedikationen und Komorbiditäten bereitstellen.

Um deren Potenzial nutzen zu können und so zur Verbesserung der Arzneimittelsicherheit beizutragen, sollen statistische Methoden weiterentwickelt werden, die sich in anderen Anwendungsgebieten bewährt haben. So steht eine große Bandbreite von Methoden für die Auswertung von Spontanmeldedaten zur Verfügung: Diese sollen zunächst umfassend verglichen und anschließend hinsichtlich ihrer Nutzbarkeit für longitudinale Daten erschlossen werden. Des Weiteren wird aufgezeigt, wie maschinelle Lernverfahren helfen könnten, seltene Risiken zu identifizieren. Zudem werden sogenannte Enrichment-Analysen eingesetzt, mit denen pharmakologische Arzneimittelgruppen und verwandte Komorbiditäten zusammengefasst werden können, um vulnerable Bevölkerungsgruppen zu identifizieren.

Insgesamt werden diese Methoden die Arzneimittelrisikoforschung anhand von GKV-Routinedaten vorantreiben, die aufgrund ihres Umfangs, der longitudinalen Erfassung sowie ihrer Aktualität eine vielversprechende Datenquelle bieten, um UAWs aufzudecken.

Schlüsselwörter

Unerwünschte Arzneimittelwirkungen · Patientensicherheit · GKV-Abrechnungsdaten · Signalerkennung · Spontanmelderegister

Detection of drug risks after approval. Methods development for the use of routine statutory health insurance data

Abstract

Adverse drug reactions are among the leading causes of death. Pharmacovigilance aims to monitor drugs after they have been released to the market in order to detect potential risks. Data sources commonly used to this end are spontaneous reports sent in by doctors or pharmaceutical companies. Reports alone are rather limited when it comes to detecting potential health risks. Routine statutory health insurance data, however, are a richer source since they not only provide a detailed picture of the patients' wellbeing over time, but also contain information on concomitant medication and comorbidities.

To take advantage of their potential and to increase drug safety, we will further develop statistical methods that have shown their merit in other fields as a source of inspiration. A plethora of methods have been proposed over the years for spontaneous reporting data: a comprehensive comparison of these methods and their potential use for longitudinal data should be explored. In addition, we show how methods from machine learning could aid in identifying rare risks. We discuss these so-called enrichment analyses and how utilizing pharmaceutical similarities between drugs and similarities between comorbidities could help to construct risk profiles of the patients prone to experience an adverse drug event.

Summarizing these methods will further push drug safety research based on healthcare claim data from German health insurances which form, due to their size, longitudinal coverage, and timeliness, an excellent basis for investigating adverse effects of drugs.

Keywords

Drug-related side effects and adverse reactions · Patient safety · Health claim data · Signal detection · Adverse drug reaction reporting systems

durchgeführt werden sollte, oder es wird eine Überwachung der gemeldeten UAW als notwendig erachtet.

Grundlegend lassen sich im Rahmen der Pharmakovigilanz als potenzielle Datenquellen die bereits angesprochenen Spontanmelderegister sowie die in den gesetzlichen Krankenkassen vorliegenden Routinedaten nennen. Der Großteil der vorgeschlagenen statistischen Methoden beruht auf Daten aus Spontanmelderegistern. Viele Verfahren basieren auf sogenannten Disproportionalitätsanalysen, bei denen aus den eingegangenen Meldungen für jede Kombination aus einem Arzneimittel und einem Ereignis (z. B. Schlaganfall) eine Vierfeldertafel (s. **Tab. 1**) erstellt wird. Für jede dieser Kombinationen wird basierend auf der entsprechenden Vierfeldertafel ein „Risikomaß" berechnet, das als Indikator für die Stärke einer potenziellen UAW dient. Zu diesen Maßen gehört z. B. das Reporting Odds Ratio (ROR; [19]), das sich als $ROR = \frac{ad}{bc}$ aus der **Tab. 1** berechnet und – grob gesprochen – die geschätzte Wahrscheinlichkeit, dass ein bestimmtes Ereignis unter Einnahme eines spezifischen Arzneimittels eintritt, mit der geschätzten Wahrscheinlichkeit vergleicht, dass dieses Ereignis unter Nichteinnahme dieses Medikaments eintritt. Dieses und andere einfache Risikomaße können bei einer sehr kleinen Anzahl an Ereignissen zu sehr hohen Werten führen, die dann fälschlicherweise als UAW angesehen würden. Solche falsch-positiven Signale können auch durch die simultane

Tab. 1 Vierfeldertafel – Anzahl der Meldungen mit einer bestimmten Kombination aus Ereignis (Ja/Nein) und Arzneimittel (Ja/Nein)

Kombination		Ereignis		Gesamt
		Ja	Nein	
Arzneimittel	Ja	a	b	a + b
	Nein	c	d	c + d
Gesamt		a + c	b + d	a + b + c + d

Überprüfung (sog. multiples Testproblem) von ggf. sehr vielen Risikomaßen entstehen. Betrachtet man etwa Kombinationen aus 1000 Arzneimitteln und 1000 Ereignissen, müssen aus einer Million Vierfeldertafeln die entsprechenden Risikomaße berechnet und inferenzstatistisch überprüft werden.

Um falsch-positive Signale zu vermeiden, wurden aufwendigere bayesianische Verfahren entwickelt, die die Risikoschätzung bei kleinen Ereignisanzahlen nach unten korrigieren (Shrinkage-Verfahren; [20–22]). Zwei der gebräuchlichsten Verfahren sind der Gamma-Poisson Shrinker [20], der bei der Food and Drug Administration (FDA) in den USA zum Einsatz kommt, und das Bayesian Confidence Propagation Neural Network (BCPNN; vgl. [21, 22]), das von der Weltgesundheitsorganisation (WHO) im Uppsala Monitoring Centre (UMC) in Schweden eingesetzt wird.

Neben diesen Ansätzen werden frequentistische Hypothesentests [19, 20, 23], penalisierte Regressionsmodelle [24] sowie Assoziationsmaße [25] und weitere bayesianische Verfahren [26, 27] in der Literatur zur Signalgenerierung diskutiert.

Um GKV-Routinedaten zu Zwecken der Pharmakovigilanz nutzen zu können, müssen geeignete Methoden für den Einsatz bei Längsschnittdaten weiterentwickelt werden. Dazu ist es sinnvoll, aus den gängigsten Methoden zur Signalgenerierung zunächst die hinsichtlich der Reduzierung falsch-positiver Signale vielversprechendsten Ansätze zu identifizieren. Zu diesem Zweck wird ein umfangreicher Methodenvergleich durchgeführt, in dem für unterschiedliche Szenarien die im Spontanmelderegister eingehenden Signale statistisch simuliert werden.

Identifikation seltener UAW mit maschinellen Lernverfahren

Neben den im vorhergehenden Abschnitt diskutierten gängigen Methoden der Pharmakovigilanz ist zu überlegen, ob die Arzneimitteltherapiesicherheit nicht auch von den in anderen Fachgebieten sehr erfolgreich eingesetzten maschinellen Lernverfahren profitieren kann. Diese dort eingesetzten Algorithmen versuchen, Muster in Lerndaten zu erkennen und mit dem generierten Wissen unbekannte Daten zu beschreiben oder Ergebnisse vorherzusagen. Es gibt eine Vielzahl maschineller Lernalgorithmen. Zwei prominente Verfahren sind Deep Learning und Random Forest.

Deep Learning ist eine Weiterentwicklung der biologisch motivierten künstlichen neuronalen Netze mit besonders vielen und neuronenreichen internen Nervenzellschichten. Die Optimierungsmethode wird u. a. sehr erfolgreich in der Sprach- [28] oder Bilderkennung [29] eingesetzt. Deep-Learning-Algorithmen haben in der Regel eine deutlich höhere Klassifikationsgenauigkeit als etablierte multivariate Klassifikationsverfahren [30] und eignen sich insbesondere dafür, nichtlineare, hochkomplexe Zusammenhänge selbstständig zu erkennen und abzubilden. Bekannte Nachteile des Deep-Learning-Ansatzes sind die hohen Anforderungen an eine performante Hardware, das Blackbox-Problem (mangelnde Möglichkeit, die inneren Abläufe und somit das Ergebnis neuronaler Netze erklären zu können) sowie die fehlende Universalität: Deep-Learning-Netze werden in der Regel problemspezifisch entwickelt und angepasst. Dafür müssen der Netzwerktyp, die Netzwerkstruktur und die Lernregel ausgewählt sowie viele weitere „Stellschrauben" adjustiert werden wie etwa die Lernrate oder das Momentum. Dieses Feintuning erfolgt durch einen Experten und eignet sich nicht für die angestrebte automatisierte Signalerkennung seltener UAW. Im Fokus steht daher die Entwicklung eines universellen Deep-Learning-Algorithmus für die Pharmakovigilanz.

Deep Learning wird auf den Routinedaten so implementiert, dass zunächst für alle Versicherten geprüft wird, ob die Entstehung bestimmter ausgewählter UAW unter Einfluss des zu untersuchenden Wirkstoffs erfolgte. Zusätzlich werden die wichtigsten verfügbaren Informationen aus dem jeweiligen Indexjahr hinzugefügt: Alter, Geschlecht, Codierung der Diagnose nach ICD (International Statistical Classification of Diseases and Related Health Problems), OPS (Operationen- und Prozedurenschlüssel), ATC (Anatomical Therapeutic Chemical Classification), EBM (Einheitlicher Bewertungsmaßstab) etc., um mögliche Confounder (Störgrößen) zu berücksichtigen. Mit all diesen Variablen wird das Deep-Learning-Netz trainiert, in neuen Daten die UAW vorherzusagen und die Klassifikationsgüte des Deep-Learning-Netzes zu bestimmen. Nach erfolgtem Lernen wird die Einnahme des zu untersuchenden Wirkstoffs bei allen Patienten auf null gesetzt („geclampt") und erneut eine Messung der Klassifikationsgüte durchgeführt. Final gelten diejenigen gelisteten UAW als potenzielle Signale, bei denen das Clamping zur stärksten Verringerung der Klassifikationsgüte führte.

Zur Identifikation von UAW kann neben Deep Learning alternativ auch ein Random Forest eingesetzt werden, bei dem es sich um ein robustes maschinelles Lernverfahren mit hoher Klassifikationsgenauigkeit handelt. Ein Vorteil gegenüber Deep Learning ist, dass zum Trainieren eines Random Forest weniger Feintuning nötig ist. Darüber hinaus ist keine Vorauswahl der möglichen UAW nötig, alle Diagnosen können gemeinsam in einem Modell analysiert werden. Hierfür wird eine sogenannte Self-controlled-Case-Series-Analyse [31] durchgeführt. Dabei werden alle Versicherten betrachtet, bei denen mindestens eine Verschreibung des interessierenden Arzneimittels vorliegt. Für alle Versicherten werden zwei Datensätze erstellt, jeweils

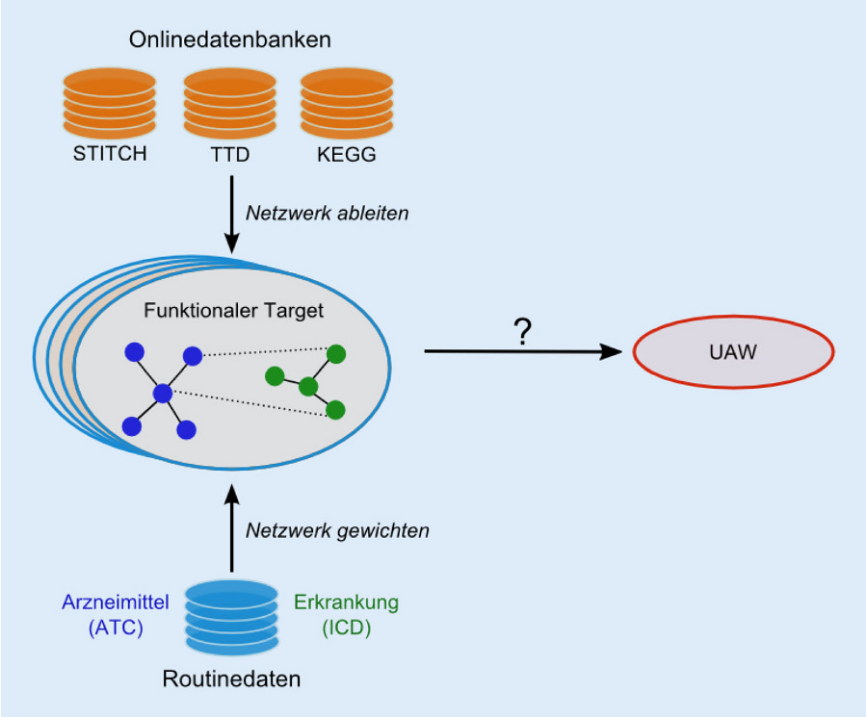


Abb. 2 ▲ Ablauf einer Enrichment-Analyse zum Auffinden funktionaler Targets (basierend auf den Routinedaten der Gesetzlichen Krankenversicherungen), die unter Arzneimittelanwendern mit Unverträglichkeiten assoziiert sind. Es werden biologische und chemische Onlinedatenbanken verwendet (z. B. Kyoto Encyclopedia of Genes and Genomes (KEGG) [40], Search Tool for Interactions of Chemicals (STITCH) [44], Therapeutic Target Database (TTD) [45]), um zu untersuchen, ob die Arzneimittel und Erkrankungen, die einem bestimmten funktionalen Target zugeordnet sind, mit der unerwünschten Arzneimittelwirkung (UAW) assoziiert sind. *ATC* Anatomisch-Therapeutisch-Chemisches Klassifikationssystem, *ICD* International Statistical Classification of Diseases and Related Health Problems

für die Zeit vor und nach der ersten Verschreibung. Für diese Datensätze werden jeweils alle Diagnosen (ICD-10) und Verschreibungen (ATC) sowie mögliche Confounder betrachtet. Auf diesen Daten wird ein Random Forest trainiert, die biaskorrigierte Variablenwichtigkeit für jede Diagnose berechnet sowie ein *p*-Wert geschätzt, basierend auf der Nullhypothese, dass keine Assoziation zwischen Diagnose und Verschreibung vorliegt. Alle signifikant assoziierten Diagnosen werden nach der Effektstärke sortiert. Die Richtungen der Effekte werden aus Vierfeldertafeln geschätzt. Das Ergebnis dieser Analyse ist eine Rangliste detektierter Signale möglicher UAW.

Beschreibung von UAW-Risikoprofilen

Eine wichtige, aber in Arzneimittelsicherheitsstudien oft nicht beantwortete Frage ist, ob bestimmte Bevölkerungsgruppen ein erhöhtes Risiko für eine spezifische UAW aufweisen. Die Verträglichkeit eines Arzneimittels ist von Person zu Person unterschiedlich und wird individuell durch verschiedene Faktoren beeinflusst, wie beispielsweise durch die genetische Ausstattung [32, 33], das Mikrobiom [34], den Lebensstil (bspw. Ernährung, Alkohol, Rauchen; [35]), durch den allgemeinen Gesundheitszustand [36] oder durch Komedikation [37]. Unterschiedliche Arzneimittel können jedoch die gleichen UAW auslösen, u. a. wenn diese einen ähnlichen unerwünschten Effekt besitzen (z. B. Hypoglykämie als Folge von unterschiedlich wirkenden Antidiabetika) oder ähnliche Wirkmechanismen aufweisen (z. B. hemmen sowohl Antidepressiva als auch Antipsychotika Muskarinrezeptoren, was zu Harnverhalt als UAW führen kann).

Um die oben erwähnten Hochrisikopatienten zu identifizieren, besteht ein mögliches Vorgehen darin, zunächst alle Versicherten, die ein bestimmtes Medikament verschrieben bekommen haben und eine spezifische UAW aufweisen, hinsichtlich ihrer Komedikation zu untersuchen. Konkret wird geprüft, ob diese Versicherten Medikamente einnehmen, die zu einer für die UAW relevanten pharmakologischen Gruppe gehören. In einem weiteren Schritt wird geprüft, ob das Kollektiv derjenigen Versicherten, die unter Einnahme des interessierenden Medikaments eine UAW erleiden, im Vergleich zu dem entsprechenden Kollektiv ohne UAW unerwartet viele weitere Risikofaktoren aufweist (engl. „enriched"). Diese Risikofaktoren werden schließlich zu einem Risikoprofil gebündelt, wodurch die einzelnen, ggf. sehr kleinen Effekte der jeweiligen Risikofaktoren kumuliert werden, was die Prognose einer UAW erleichtert.

Diese Bündelung von Risikofaktoren wird häufig auch als „Enrichment" bezeichnet, wobei der Begriff „Enrichment-Analyse" zwar aus der genetischen Forschung stammt, das Prinzip aber auch in der Pharmakologie angewandt wird, um UAW basierend auf molekularen Targets vorherzusagen [38, 39]. Diese Targets können durch unterschiedliche Arzneimittel aktiviert werden, die eine hohe pharmakologische Ähnlichkeit aufweisen. Das gilt besonders für Arzneimittel derselben Wirkstoffklasse.

In der genetischen Epidemiologie bündeln Gene-Set-Enrichment-Analysen (GSEA) Gene oder Proteine in funktionale Gruppen und untersuchen, welche dieser Gruppen mit der Erkrankung assoziiert sind. Die Einordnung von Genen in solche funktionalen Gruppen erfolgt mittels entsprechender Onlinedatenbanken (z. B. KEGG [40], GO [41]). Mittlerweile gibt es vielfältige Strategien für GSEA [42], u. a. auch topologiebasierte Verfahren. Diese berücksichtigen zusätzlich, inwiefern Gene aus einer funktionalen Gruppe gemeinsam exprimiert werden [43]. Der Vorteil von GSEA ist, dass sie Einblicke in den biologischen Kontext multipler genetischer Risikofaktoren bieten und damit auch Ideen für Krankheitsmechanismen und mögliche Behandlungsansätze liefern können. Allerdings sind die Ergebnisse stark abhängig von der Definition

der Gengruppen und werden daher als hypothesengenerierend verstanden.

Bei der Charakterisierung von UAW-Risikoprofilen sollen Enrichment-Analysen helfen, die pharmakologisch relevanten Gruppen zu identifizieren, die bei Patienten, die ein bestimmtes Arzneimittel einnehmen, mit einer UAW assoziiert sind. Dafür werden Arzneimittel und Erkrankungen in funktionale Targets (z. B. Rezeptoren, Enzyme, molekulare Prozesse, Wirkstoffklassen, höhere Ebenen der ICD-Codierung) anhand der einschlägigen Onlinedatenbanken (z. B. KEGG Drug, STITCH [44], TTD [45], ChEMBL [46]) eingruppiert. Dabei kann die pharmakologische Ähnlichkeit von Arzneimitteln innerhalb eines Targets bspw. anhand des Chemical-Similarity-Scores [47] bewertet werden und die Information, welche Erkrankungen häufiger miteinander auftreten, durch den Cormobidity-Score [48] berücksichtigt werden. Die Anwendung der Enrichment-Analyse in der Pharmakovigilanz wird in **Abb. 2** illustriert.

Mit einer auf Routinedaten zugeschnittenen Enrichment-Methode könnten daher die funktionalen Targets identifiziert werden, die ein erhöhtes UAW-Risiko aufweisen. Das hilft, einerseits den biologischen Mechanismus hinter der UAW zu erklären und andererseits Risikoprofile für vulnerable Bevölkerungsgruppen zu erstellen, die in diesen funktionalen Targets mit entsprechenden Komedikationen oder Komorbiditäten „enriched" sind.

Diskussion

Die Nutzung von Versichertendaten für die Pharmakovigilanzforschung erscheint äußerst vielversprechend, beinhaltet aber auch einige Herausforderungen, insbesondere da die gesammelten Routinedaten der GKVen nicht für Forschungszwecke, sondern für die Abrechnung erbrachter Leistungen im Gesundheitswesen erhoben werden. Die sich daraus ergebenden Unschärfen und mögliche Fehler in den Daten können in konfirmatorischen pharmakoepidemiologischen Studien basierend auf solchen Datenbanken z. B. durch ein geeignetes Studiendesign und entsprechende Methoden berücksichtigt werden. Will man diese Daten, wie oben beschrieben, für automatisierte Signalgenerierungsstudien nutzen, ist dies nicht zu leisten.

Hier müssen andere Lösungen wie die vorgestellten automatisierten Lernverfahren gefunden werden. Dabei greifen die obigen methodischen Weiterentwicklungen insbesondere drei wesentliche Probleme der Arzneimittelsicherheitsforschung auf: (1) Falsch-positive Signale können zu einer Verunsicherung von Patienten, aber auch von Ärzten führen und somit eine adäquate Versorgung gefährden. (2) Häufig reichen die Fallzahlen in pharmakoepidemiologischen Studien nicht aus, um auch seltene Ereignisse mit einer vorgegebenen statistischen Sicherheit zu erkennen, was zu einer falschen Einschätzung des Gefährdungspotenzials durch ein Arzneimittel führen kann. (3) Es ist bekannt, dass Arzneimittelrisiken nicht in gleichem Maße bei jedem Patienten auftreten. Dennoch können identifizierte Risiken zu einer Marktrücknahme führen, die für Patienten, bei denen das Medikament keine Schäden hervorgerufen hat, eine schlechtere Versorgung zur Folge hat. Damit ist die Erkennung von Risikoprofilen essenziell für eine bessere Einschätzung des Gefährdungspotenzials von Arzneimitteln.

Nicht zuletzt bedeutet eine systematischere Erfassung von potenziellen UAW in Anbetracht der Datenmenge eine große Zeit- und Kostenersparnis im Gesundheitswesen. Durch Automatisierungsprozesse in der Datenverarbeitung könnte die Effizienz der Signalgenerierung in Zukunft gesteigert werden.

Fazit

Aufgrund des Umfangs, der Kontinuität und der standardisierten Erfassung erscheint die Nutzung von GKV-Routinedaten für die Pharmakovigilanzforschung zur Verbesserung der Qualität der Pharmakotherapie überaus attraktiv. Dabei kommt der Erkennung von falsch-positiven Signalen und seltenen UAW sowie der Ermittlung von spezifischen Risikoprofilen eine besondere Bedeutung zu. Auch für zukünftige Fragestellungen dürften Routinedaten noch viel Potenzial bieten.

Korrespondenzadresse

Dr. R. Foraita
Leibniz-Institut für Präventionsforschung und Epidemiologie – BIPS
Achterstr. 30, 28359 Bremen, Deutschland
foraita@leibniz-bips.de

Danksagung. Dieser Artikel entstand als Teil des Projekts „Nutzung von Routinedaten zur Pharmakovigilanz in Deutschland: Methodenentwicklung und erste Anwendungen", kurz PV-Monitor, das im Rahmen des Innovationsfonds des Gemeinsamen Bundesausschusses unter dem Förderkennzeichen 01VSF16020 gefördert wird.

Einhaltung ethischer Richtlinien

Interessenkonflikt. R. Foraita, L. Dijkstra, F. Falkenberg, M. Garling, R. Linder, R. Pflock, M. R. Rizkallah, M. Schwaninger, M. N. Wright und I. Pigeot geben an, dass kein Interessenkonflikt besteht.

Dieser Beitrag beinhaltet keine von den Autoren durchgeführten Studien an Menschen oder Tieren.

Literatur

1. Lazarou J, Pomeranz BH, Corey PN (1998) Incidence of adverse drug reactions in hos-pitalized patients: a meta-analysis of prospective studies. JAMA 279:1200–1205
2. European Commission (2008) Proposal for a regulation amending, as regards pharmacovigilance of medicinal products for human use. Regulation (EC) No 726/2004. http://ec.europa.eu/health/files/pharmacos/pharmpack_12_2008/pharmacovigilance-ia-vol1_en.pdf. Zugegriffen: 12. Jan. 2018
3. Oscanoa TJ, Lizaraso F, Carvajal A (2017) Hospital admissions due to adverse drug reactions in the elderly. A meta-analysis. Eur J Clin Pharmacol 73:759–770
4. Bouvy JC, De Bruin ML, Koopmanschap MA (2015) Epidemiology of adverse drug reactions in Europe: a review of recent observational studies. Drug Saf 38:437–453
5. Stausberg J, Hasford J (2011) Drug-related admissions and hospital-acquired adverse drug events in Germany: a longitudinal analysis from 2003 to 2007 of ICD-10-coded routine data. BMC Health Serv Res 11:134
6. Graham DJ, Campen D, Hui R et al (2005) Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. Lancet 365:475–481
7. Sawicki PT, Bender R, Selke GW, Klauber J, Gutschmidt S (2006) Assessment of the number of cardio- and cerebrovascular events due to rofecoxib (Vioxx) in Germany between 2001 and 2004. Med Klin (Munich) 101:191–197
8. Bundesärztekammer (2015) (Muster-)Berufsordnung für die in Deutschland tätigen Ärztinnen

und Ärzte in der Fassung des Beschlusses des 118. Deutschen Ärztetages 2015 in Frankfurt am Main. Dtsch Arztebl Int 112:1348 ((A3, § 6))
9. Arzneimittelkommission der deutschen Ärzteschaft (2016) Was geschieht mit den Meldungen an die AkdÄ? https://www.akdae.de/Arzneimittelsicherheit/UAW-Meldung/Info/UAW-Meldung-Analyse.html. Zugegriffen: 26. Jan. 2018
10. European Medicines Agency (2012) Europäische Datenbank gemeldeter Verdachtsfälle von Arzneimittelnebenwirkungen: Hintergrund. http://www.adrreports.eu/de/background.html. Zugegriffen: 26. Jan. 2018
11. Suling M, Pigeot I (2012) Signal detection and monitoring based on longitudinal healthcare data. Pharmaceutics 4:607–640
12. Pigeot I, Windeler J (2005) Klinische Prüfung nach der Zulassung. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 48:580–585
13. Stephenson WP, Hauben M (2007) Data mining for signals in spontaneous reporting databases: proceed with caution. Pharmacoepidemiol Drug Saf 16:359–365
14. Goldman S (1998) Limitations and strengths of spontaneous reports data. Clin Ther 20(Suppl C):C40–C44
15. Bates D, Evans R, Murff H, Stetson P, Pizziferri L, Hripcsak G (2003) Detecting adverse events using information technology. J Am Med Inform Assoc 10:115–128
16. Harpaz R, DuMouchel W, Shah N, Madigan D, Ryan P, Friedman C (2012) Novel data-mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Ther 91:1010–1021
17. Garbe E, Pigeot I (2015) Der Nutzen großer Gesundheitsdatenbanken für die Arzneimittelrisikoforschung. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 58:829–837
18. Pigeot I, Ahrens W (2008) Establishment of a pharmacoepidemiological database in Germany: methodological potential, scientific value and practical limitations. Pharmaco-epidemiol. Drug Saf 17:215–223
19. Van Puijenbroek EP, Bate A, Leufkens HGM, Lindquist M, Orre R, Egberts ACG (2002) A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf 11:3–10
20. DuMouchel W (1999) Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat 53:177–190
21. Bate A, Lindquist M, Edwards IR et al (1998) A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol 54:315–321
22. Norén GN, Bate A, Orre R, Edwards IR (2006) Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. Stat Med 25:3740–3757
23. Ahmed I, Dalmasso C, Haramburu F, Thiessard F, Broët P, Tubert-Bitter P (2010) False discovery rate estimation for frequentist pharmacovigilance signal detection methods. Biometrics 66:301–309
24. Caster O, Madigan D, Norén GN, Bate A (2008) Large-scale regression-based pattern discovery in international adverse drug reaction surveillance. Proceedings of the KDD-08 Workshop on Mining Medical Data, S 24–27
25. Roux E, Thiessard F, Fourrier A, Be B (2005) Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance. Ieee J Biomed Health Inform 9:518–527
26. Ahmed I, Haramburu F, Fourrier-Réglat A et al (2009) Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting. Stat Med 28:1774–1792
27. Madigan D, Ryan P, Simpson S, Zorych I (2010) Bayesian methods in pharmacovigilance. Bayesian Stat 9:421–438. https://doi.org/10.1093/acprof:oso/9780199694587.001.0001
28. Mohamed AR, Sainath TN, Dahl G, Ramabhadran B, Hinton GE, Picheny MA (2011) Deep belief networks using discriminative features for phone recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, 2011, S 5060–5063
29. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
30. Linder R (2006) Lernstrategien zur automatisierten Anwendung künstlicher neuronaler Netzwerke in der Medizin. Logos-Verlag, Berlin
31. Whitaker HJ, Farrington CP, Spiessens B, Musonda P (2006) Tutorial in biostatistics: The self-controlled case series method. Stat Med 25:1768–1797
32. Meyer UA (2000) Pharmacogenetics and adverse drug reactions. Lancet 356:1667–1671
33. Phillips KA, Veenstra DL, Oren E, Lee JK, Sadee W (2001) Potential role of pharmacogenomics in reducing adverse drug reactions: a systematic review. JAMA 286:2270–2279
34. Rizkallah MR, Saad R, Aziz RK (2010) The Human Microbiome Project, personalized medicine and the birth of pharmacomicrobiomics. Curr Pharmacogenomics Person Med 8:182–193
35. Alomar MJ (2014) Factors affecting the development of adverse drug reactions. Saudi Pharm J 22:83–94
36. Dumbreck S, Flynn A, Nairn M et al (2015) Drug-disease and drug-drug interactions: systematic examination of recommendations in 12 UK national clinical guidelines. BMJ 350:h949
37. Stewart D, Gibson-Smith K, MacLure K et al (2017) A modified Delphi study to determine the level of consensus across the European Union on the structures, processes and desired outcomes of the management of polypharmacy in older people. PLoS ONE 12:e188348
38. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK (2007) Relating protein pharmacology by ligand chemistry. Nat Biotechnol 25:197–206
39. Lounkine E, Keiser MJ, Whitebread S et al (2012) Large-scale prediction and testing of drug activity on side-effect targets. Nature 486:361–367
40. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45:D353–D361
41. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25(1):25–29
42. Mooney MA, Wilmot B (2015) Gene set analysis: a step-by-step guide. Am J Med Genet B Neuropsychiatr Genet 168:517–527
43. Wang Q, Yu H, Zhao Z, Jia P (2015) EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression profiles. Bioinformatics 31(15):2591–2594. https://doi.org/10.1093/bioinformatics/btv150
44. Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M (2016) STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res 44:D380–D384
45. Li YH, Yu CY, Li XX et al (2018) Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. Nucleic Acids Res 46:D1121–D1127
46. Gaulton A, Hersey A, Nowotka M et al (2017) The ChEMBL database in 2017. Nucleic Acids Res 45:D945–D954
47. Lo Y-C, Torres JZ (2016) Chemical similarity networks for drug discovery. In: Chen T (Hrsg) special topics in drug discovery. Intech. https://www.intechopen.com/books/special-topics-in-drug-discovery/chemical-similarity-networks-for-drug-discovery. Zugegriffen: 30. Jan. 2018
48. Hude Q, Vijaya S, Patricia H et al (2005) Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 43:1130–1139

B.3 Predicting patient risk for adverse drug reactions in health care claims data using functional targets

Contribution to the manuscript: R. Foraita, I and L. Dijkstra planned the concept. I conducted the literature research, data curation and transformation pipeline. Statistical analysis plan was compiled by R. Foraita and I in collaboration with L. Dijkstra, I. Pigeot and AFX. Wilhelm. I optimized the methods performance with R. Foraita and L. Dijkstra. I wrote the initial draft, R. Foraita revised parts of the draft.

Predicting Patient Risk for Adverse Drug Reactions in Health Care Claims Data using Functional Targets

Running title: Using biological functional targets for adverse drug reaction prediction in health care claims data: Application on gastrointestinal and intracranial bleeding

Initial authors: Mariam R. Rizkallah[1,2] & Ronja Foraita[1*]

[1] Leibniz Institute for Prevention Research and Epidemiology − BIPS, Achterstraße 30, 28359 Bremen, Germany
[2] Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany

[*] Corresponding Author:
Ronja Foraita
Achterstraße 30
28359 Bremen
Germany
Tel.: +49 421 218 56954
Fax: +49 421 218 56941
E-Mail: foraita@leibniz-bips.de

Keywords: comorbidities, health insurance, pharmacovigilance, prediction, risk factors

Declarations:

Funding: This work is funded by the innovation fund ("Innovationsfonds") of the Federal Joint Committee in Germany (grant number: 01VSF16020).

Conflicts of interest: The authors declare no conflicts of interest.

Authors' contributions: MRR and RF conceptualized the study, MRR compiled the data, MRR and RF optimized the statistical methods. MRR wrote the initial draft, RF revised parts of the draft.

Statement: The content of this paper has not been presented or published prior to this submission.

Journal: *Drug Safety* https://www.springer.com/journal/40264/submission-guidelines

Abstract

Word count: 248 words

Background: Adverse drug reactions (ADRs) represent a burden on health care systems. Identifying populations at increased risk based on co-administered drugs and co-morbidities requires ADR prediction and risk factor identification from comprehensive sources such as health care claims databases. We present a strategy for predicting ADRs, grouping drug and disease predictors according to their biological functional targets (FTs), basing ADR prediction on group-ADR associations. Exploiting domain knowledge may better explain predictors relationships and increase predictive power.

Methods: We compared three settings: grouping according to FTs or WHO drug/disease classification, and no grouping, applying: random forests (RF) and block forests (BF), LASSO, LASSO for a constructed group variable (NGL), and an extension of the adaptive rank truncated product (ARTP). We used the German Pharmacoepidemiological Research Database to construct two matched case-control studies for gastrointestinal bleeding (GIB) and intracranial bleeding (ICB). We controlled for age, sex, region of residence and time-to-event. FT information were curated from the Therapeutic Target Database.

Results: In both samples, GIB (N=64,720) and ICB (N=34,600), LASSO, RF and BF (FT-grouping) performed best. In the ICB sample, NGL (WHO-grouping) performed comparable to the LASSO and RF. The ARTP performed poor showing slight improvement using WHO-grouping.

Conclusion: BF using FTs is a candidate method for risk prediction. Further investigation is required to determine the effect of data set size, group structure and size on performance. This strategy is expendable with drug-target score data and potentially dosage information.

1 Introduction

Adverse drug reactions (ADRs) represent a burden on the health care system as they lead to patient morbidity and mortality. In Europe alone, 3-10% of all hospital admissions are due to ADRs [1]. A patient's response to a drug, including susceptibility to ADRs, is the sum of many factors such as: genetic makeup [2, 3], microbiome [4], lifestyle, e.g., nutrition, alcohol consumption and smoking [5], co-morbidities [6], and concomitant drug use [7]. Particularly in elderly patients, polypharmacy and multi-morbidity can lead to an increased risk of ADRs [6, 8].

Identifying groups of patients at increased risk of ADRs becomes of utmost importance. Such identification can be achieved based on a number of patient characteristics, particularly co-administered drugs and co-morbidities. Therefore, the incorporation of various data types at high-resolution on a large scale is required. A comprehensive resource of such data is health care claims data. Health care claims databases store routinely collected data for reimbursement purposes by statutory health insurances (SHIs). They contain demographic information (e.g., age, sex, occupation), prescription information (e.g., drug name, dose, duration, and possibly route of administration and therapeutic indication), and diagnosis information (e.g., in- and outpatient diagnoses and procedures [9, 10, 11]. The quality, magnitude and comprehensiveness of their data qualify health care claims databases as a good source for ADR prediction.

In Figure 1a, the classical approach to predict patient risk is schematically represented. Classically, patient risk is predicted based on the associations between individual risk factors (drugs and diseases) and the ADR. This approach is limited by 1) the restricted available information for patients exposed to these drugs, and 2) the scale at which statistical models can handle and utilize such a relatively large number of variables. Methods for variable selection (e.g., penalized logistic regression) are important data dimensionality reduction approaches for large-scale signal detection studies. Another approach to reduce data dimensionality is testing for association between a group of covariates (e.g., drugs/diseases) and a specific outcome (e.g., ADR). This concept is well-established in genetic epidemiology (shown schematically in Figure 1b). In genetic epidemiology, pathway analysis approaches allow for combining evidence for associations between single covariates (e.g., genes) and the outcome (e.g., phenotype), which 1) leads to better signal detection, and 2) helps to interpret the risk factors according to their biological pathways.

Here, we propose that instead of assessing the associations between the drugs and diseases, and the ADR directly, the associations between the groups and the ADR shown should be investigated. Further, as sketched in Figure 2, we propose that drugs and diseases are grouped by the functional targets (FTs) they interact with. We define a FT as: a pathway of interacting biomolecules (e.g., enzymes, receptors) that are affected by the drug [12] or associated with a disease. We hypothesize that drugs and diseases involved with an FT are more likely to lead to the ADRs associated with that FT.

There are several public repositories of largely manually curated biological and chemical databases that link drugs and diseases to FTs, e.g., the Therapeutic Targets Database (TTD) [13], ChEMBL [14], Search Tool for Interacting Chemicals (STITCH) [15], and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [16]. To the best of our knowledge, TTD is the most comprehensive publicly available online database that allows linking drugs and

diseases to the same FT. First, TTD provides clear, comprehensive information on drug-disease relationships according to the ICD coding system. KEGG, in contrast, is highly selective as it provides information on the diseases that primarily have an underlying genetic cause. Second, TTD can be readily cross-referenced with the well-known knowledge base KEGG, and with ICD and ATC systems (used by SHIs data). Therefore, we considered using TTD for curating grouping structures to annotate the predictors.

Here, we elaborate on the proposed approach using an example of the direct oral anticoagulants (DOACs). The four DOACs, dabigatran, rivaroxaban, apixaban and edoxaban, have been used for thromboembolism prevention and/or treatment, acting via direct inhibition of coagulation cascade factors (i.e., enzymes) [17]. In particular, the target of dabigatran is thrombin, while the target of the other three DOACs is the coagulation factor Xa. Those two enzymes are encoded by genes in the complement and coagulation cascade pathway (KEGG ID: hsa04610). Therefore, in our approach, DOACs belong to the functional target-based group hsa04610, because their target enzymes are encoded by genes in this pathway. Furthermore, underlying co-morbidities can affect drug choice and ADRs; selective serotonin reuptake inhibitors would increase the risk of bleeding in depressed patients with myocardial infarction [6]. A FT-based view can explain ADRs, providing an understanding of the combined drug-disease effect. By exploiting domain knowledge to assign drugs/diseases to groups, we can possibly improve risk prediction by increasing the power for detecting associations. In addition, pooling the data within each group reduces data dimensionality. Moreover, target-based prediction of ADRs might help resolve target pathways (more importantly unintended target pathways) of drugs, which can better explain the underlying mechanisms of ADRs.

Various methods were developed integrating domain knowledge on drug and disease molecular targets, and a few were applied to various forms of electronic health care data with the purpose of ADR prediction. Examples of the developed methods include data mining and/or machine learning methods using information on molecular similarity between drugs [18], drug molecular pathways [19], and drug-drug interactions [20], or using ATC classes [21, 22]. Concerning health care claims databases, data mining methods have been used for safety signal detection (reviewed in [23]). Nevertheless, to the best of our knowledge, such approach is yet to be used for large-scale ADR prediction using health care claims data.

In this study, we aim at comparing the predictability of an event of interest (EI) in patients given his/her drug exposures and diseases using FTs. We consider statistical methods that are able to exploit this underlying group structure, some of which were able to infer the FTs (and drugs and diseases) that are associated with the risk of the event. We compare the effect of FTs grouping structures to that of the WHO classification systems for drugs (Anatomical Therapeutic Chemical Classification System; ATC) and diseases (International Classification of Diseases; ICD). We apply our approach to health care claims data from the German Pharmacoepidemiological Research Database (GePaRD) [24] in two case-control studies considering two adverse drug events: gastrointestinal bleeding (GIB) and intracranial bleeding (ICB), both are linked to the use of DOACs.

2 Methods

2.1 Data source and data privacy

The data source for this study is GePaRD [24] established by the Leibniz Institute for Prevention Research and Epidemiology—BIPS. GePaRD is based on claims data from four statutory health insurance (SHI) providers in Germany, and currently includes information on ~25 million persons who have been insured with one of the participating providers since 2004 or later. In addition to demographic data, GePaRD contains information on drug dispensations as well as outpatient (i.e., from general practitioners and specialists) and inpatient services and diagnoses. Per data year, there is information on approximately 20% of the general population and all geographical regions of Germany are represented. Diagnoses are validated according to the International Classification of Diseases, 10th revision, German Modification (ICD-10-GM). Drug dispensations are mapped according to the Anatomical Therapeutic Chemical Classification System (ATC).

In Germany, the utilization of health insurance data for scientific research is regulated by the Code of Social Law. All involved health insurance providers as well as the German Federal Office for Social Security and the Senator for Health, Women and Consumer Protection in Bremen as their responsible authorities approved the use of GePaRD data for this study. Informed consent for studies based on claims data is required by law unless obtaining consent appears unacceptable and would bias results, which was the case in this study. According to the Ethics Committee of the University of Bremen, studies based on GePaRD are exempt from institutional review board review.

2.2 Study population and design

We conducted two case-control studies nested in a cohort of insurants who were required to: 1) be continuously insured from July 1, 2014 until December 31, 2016 with no occurrences of the event of interest (EI) until March 31, 2015, and 2) to have complete demographic information and had to be born not earlier than 1997. Cohort entry was January 1, 2015. Cohort exit was the first of the following dates: onset of the EI, death or end of study period (December 31, 2016).

Cases were defined as the patients who were hospitalized for either gastrointestinal bleeding (GIB; first case-control study) or intracranial bleeding (ICB; second case-control study). GIB ICD-10-GM codes are: I983, K226, K228, K2280, K2281, K2288, K250, K252, K254, K256, K260, K262, K264, K266, K270, K272, K274, K276, K280, K282, K284, K286, K290, K3182, K5522, K5532, K5582, K5701, K5703, K5711, K5713, K5721, K5723, K5731, K5733, K5741, K5743, K5751, K5753, K5781, K5783, K5791, K5793, K625, K661, K920, K921, and K922. ICB ICD-10-GM codes are: I61, I610, I611, I612, I613, I614, I615, I616, I618, I619, I60, I600, I601, I602, I603, I604, I605, I606, I607, I608, I609, I62, I620, I6200, I6201, I6202, I6209, I621, I629, S0633, S0634, S064, S065, and S066. ICD code descriptions are in SI_File1. The admission date is referred to as index date. Four controls were matched to each case with respect to sex, year of birth and index date. Each control was assigned an index date that resulted in the same follow-up time as for the corresponding case. Cases of an EI were not eligible to be selected as controls for the same EI; controls of one EI were eligible to be selected as controls for the other EI.

2.3 Predictor assessment

The following categories of potential predictors were considered in the analysis: disease diagnoses, drug dispensations, demographic variables (sex, age, region code). Disease predictors were obtained from in- and outpatient diagnosis data (i.e., main discharge diagnosis, secondary and auxiliary diagnosis, diagnosis for ambulatory treatment, or hospitalization diagnosis) prior to the onset of the EI as ICD-10-GM codes. Drug predictors were obtained from reimbursable dispensation data prior to the onset of the EI as ATC codes. Demographic data and time-to-event (i.e., the number of days between the study entry and the onset of EI in days) were treated as adjustment variables.

2.4 Annotation of predictors into group structures

Two annotation schemes were considered for grouping drugs and diagnoses: classical drug and disease classification systems by the WHO, and functional target-based annotation. First, for the WHO grouping, we aggregated predictors according to the WHO respective drug and disease classification systems. For dispensation data, codes were truncated to ATC 4th digit, while disease codes were truncated to ICD 3rd digit. Second, groups of predictors belonging to a functional target were aggregated according to drug and disease target information as per Therapeutic Target Database (TTD, Update: 6.1.01) [25].

We used the following data from TTD: drug-target gene data and disease-target gene data. A target gene is a gene coding for, for most drugs, a molecule that the drug interacts with to exert its effect, or, in case of a disease, a molecule linked to the disease (e.g., coagulation factor Xa and atrial fibrillation, bleeding). Predictors belonging to a target gene were further aggregated and mapped to human KEGG pathways as: drug-pathway and disease-pathway relationships. A pathway consists of the genes coding for the interacting biomolecules that are affected by the drug or associated with a disease. Pathway-based aggregation is intended to reduce data dimensionality and provide a pathway-centered interpretation.

For each grouping scheme, there is a possibility that a predictor cannot be assigned to a group, for example, because the target gene of a drug or a disease is not yet discovered or curated. This mainly applies to FT grouping, as in WHO grouping, a drug or disease, respectively, hierarchically belongs to an ATC or an ICD group. In case of singletons (predictors without FT information), two strategies were followed: 1) creating a group of singletons (i.e., grp262, denoted as 1gp), and 2) splitting singletons into 1-member groups (1:n_singletons; denoted as split).

2.6 Statistical analyses

Four non-group-based and group-based statistical methods were compared in their ability to predict the risk of GIB and ICB using health care claims data. We fitted a prediction model for each EI and for each method based on the grouped or single predictors. Both case-control samples were split so that cases with an even ID number were used as to fit the model and cases with an odd ID number were used to evaluate the prediction performance.

The methods applied were chosen with regard to their ability to analyze and group high-dimensional data. Additionally, we applied those methods that were implemented in R (version 4.0.2) and able to process the data in reasonable time. The chosen methods were the LASSO [26, 27], random forest (RF) [28], block forest (BF) [29],

the LASSO for a constructed group variable (i.e., naïve-group LASSO; NGL), and the adaptive combination of rank truncated product (ARTP) [30].

First, we considered regularized regression methods because they exploit sparsity and they are able to detect signals in high-dimensional data sets. This is applicable here, as we hypothesize that only a minor proportion of the marketed drugs and diagnoses could cause the EI. The most popular penalized regression method, the LASSO, was applied for single and grouped predictors as implemented in the R package glmnet (v 4.0.2) [31]. To apply the LASSO with grouped predictors (i.e., NGL), we constructed a group variable based on the sum of predictors within one group (i.e., FT-based or WHO classification-based), multiplied by $1/\sqrt{n}$ where n is the number of predictors within group, to account for the varying group sizes. We utilized ten-fold cross-validation to select the penalty term λ.

Second, regarding machine learning approaches, we considered two methods that could analyze ungrouped and grouped high-dimensional data, which are, respectively, RF and BF. BF is a further development of RF that is able to combine different types of omics data for outcome prediction. While RF is known to capture complex dependence structures in data, BF additionally allows for including *a priori* known group structures to improve prediction performance. To estimate RFs, we used the function ranger (v 0.12.1) [32] with its default settings, where the variable importance was determined by permutations. We also used the package blockForest (v 0.2.4) to estimate the forests using the same settings as for RF, while adjusting the number of groups of tuning parameter values to 50 (for computational resources and time constraints), and the number of trees in each forest during tuning parameter optimization to 50.

Third, the ARTP was considered, which is a gene set enrichment method that was originally designed for single nucleotide polymorphism (SNP) data [30]. It is a hypothesis testing approach to select biological pathways that are enriched with genetic variants to be associated with a phenotype. The method preserves the correlation structure between genes by using permutation tests, and it has the potential to detect subtle effects of genetic variants in a pathway that might be missed when assessed individually. The ARTP uses p values from any statistical association test performed between individual SNPs and the disease outcome. We adopted the ARTP method for detecting associations between the EI and the groups when using binary health care claims data, and modified for individual risk predictions. We implemented it in R as the ARTPredict package [33]. The ARTP used p values from an adjusted logistic regression model, and used permutation tests (n = 50) based on the p values preserving group correlation structure. It then used a cutoff for group selection p value ≤ 0.05.

In summary, the predictors were analyzed according to either: no grouping (ng), FT-grouping (FT-g), or grouping according to the WHO drug/disease classification (WHO-g). For each EI, the methods were applied in the following settings: RF for ng, BF for FT-g and WHO-g, the LASSO for ng, the NGL for FT-g and WHO-g, and the ARTP for FT-g and WHO-g. The LASSO and NGL models were adjusted for sex, age, GKZ5 (truncated at the 3rd character), and time-to-event. For all grouping settings, the two approaches for handling predictors with unknown FT (i.e., singletons) were applied (i.e., 1gp and split) in case of NGL and the ARTP. For BF, which is designed to handle blocks of omics data, we only used the 1gp approach. For all these methods, non-informative variables (i.e., no variance variables) were excluded. An ensemble prediction was evaluated as well; a case is predicted when >

50% of the methods predict it as a case. Ensemble prediction was compared when using: 1) FT-g (1gp) + ng, 2) WHO-g (1gp) + ng, and 3) FT-g (1gp) + WHO-g (1gp) + ng.

The performance of the statistical methods was assessed with respect to accuracy, precision, recall, the area under the curve, and the area under the precision-recall curve (PR-AUC) of the methods in predicting the selected EI in the test dataset. We considered the PR curve for prediction evaluation, as it is argued to be more informative than the receiver operating characteristics (ROC) curve in case of evaluating binary classifiers on imbalanced datasets [34]. We calculated the baseline value for a random classifier for the PRC-AUC as follows: baseline performance = number of positives / (number of positives + number of negatives). Associated variables or groups are selected based on either variable coefficient (LASSO), variable importance (RF), constructed group variable coefficient (NGL), block split value (BF), or block *p* value (ARTP).

3 Results

3.1 Study population description

The cohort included 7,140,746 persons; 12,944 cases of GIB and 6,920 cases of ICB were identified to whom we matched 51,776 and 27,680 controls, respectively. Figure 3 illustrates the study flowchart. Seventy-nine insurants were considered cases in both subcohorts. In both nested case-controls samples, the majority of cases were male (GIB: 60 %, ICB: 62.4%) and the mean age was 66 (GIB) and 67 (ICB) years [standard deviations (SD) (GIB: 17, ICB: 16) see Table 1 and SI_File2]. The mean number of days in the sample until the event occurred was 417 for GIB (SD: 185) and 411 for ICB (SD: 185). Death within the study period occurred in 2.8% of GIB cases and 8% of ICB cases, while not among controls. Moreover, we examined the proportion of patients with DOAC dispensations in the study period, those were 6.1% of the total GIB sample and 5% of the ICB cohort.

The number of drug and disease predictors were 8,577 (GIB) and 7,847 (ICB); the majority of which are diseases (see Table 1 and SI_File2). In both samples, the largest proportion of predictors could be grouped using the FT-grouping (82%). Furthermore, the frequencies of group sizes are presented in Figure 4. In both grouping schemes and both samples, the larger the group size, the less abundant they are in the samples. Concerning FT-grouping, disregarding singletons, the group sizes range from 2 to 4000 predictors (4762 in GIB and 4310 in ICB), with the most occurring group sizes falling into the 2-402 predictor/group category, while only a small proportion of groups have sizes larger than 3000 predict/group. Concerning WHO-grouping, the group sizes range from 1 to 36 predictors in GIB and 1 to 29 predictors in ICB), with the largest proportion of group sizes falling into the 1-6 followed by 6-11 predictor/group categories.

3.2 Outcome prediction

The predictive performance of each of the statistical methods applied was evaluated according to standard classification measures. Table 2 shows the area under the precision-recall curve, while the full performance spectrum is presented in SI_File4. LASSO, RF and BF using FT-grouping performed best in both case-control samples. The PR-AUC is comparable for these three methods in the GIB sample (0.7). In the ICB sample, the NGL using WHO-grouping predicted the outcome as good as the LASSO or RF (0.8), while BF outperformed all these

methods when FT-grouping was used (0.825). The prediction performance of the ARTP was in general poor, nevertheless, when using WHO-grouping, the performance slightly improved (0.5 *vs.* 0.45 in GIB and 0.6 *vs.* 0.3 in ICB). Finally, we assessed the performance of an ensemble prediction based on best FT-g predictions (where singleton predictors are treated as 1gp), WHO-g or both, which were all comparable in both samples (0.71 in GIB and 0.82 in ICB).

3.3 Selected predictors

Here we present a quantification of top selected predictors and groups by the methods compared in this study. The number of selected predictors by the LASSO and RF, and of the selected groups by NGL, BF and ARTP for GIB and ICB are presented in Table 3, while the importance values of the selected variables and groups are presented in Table 4.

In general, a consistently larger number of predictors were selected in the GIB study compared to the ICB study, except for NGL FT-g 1gp (187 in GIB and 192 in ICB). The largest number of predictor groups were selected in case of WHO-g grouping by the NGL (487 in GIB and 359 in ICB), and by the ARTP (871 in GIB and 523 in ICB). Concerning predictor variables, the LASSO selected 840 variables for GIB, compared to 107 for ICB. The lowest number of predictor groups were selected by NGL FT-g split in ICB (52). We also assessed the number of combined selected variables by ensemble of methods either for FT-g only, WHO-g only or both, where FT-g selected fewer variables and groups (1,158 in GIB and 437 in ICB) than WHO-g (1,965 in GIB and 899 in ICB) scheme in both studies. It is important to notice for the overlap between the groups from each grouping scheme in the ensemble of potential risk factors (2,259 in GIB and 1,187 in ICB).

In Table 4, the list of the highest ranked variables and groups (top 10) are presented, as by RF, BF using FT grouping and WHO grouping. In the GIB sample, RF ranked highest the ICD codes that are either included in case definition and or are directly linked to the EI (e.g., K92, D500, D62), while BF FT-g and WHO-g ranked highest the ICD groups and FTs that are involved in cancer and urinary system disorders (e.g., hsa05219, hsa04115, hsa05212, D41, C24). Concerning cardiovascular-relevant predictors, BF FT-g ranked highest complement and coagulation cascades (hsa04610) and renin-angiotensin system (hsa04614) pathways, while BF WHO-g ranked highest cardiomyopathy (I43).

In the ICB sample, RF ranked highest the ICD codes that are either included in case definition and or are directly linked to the EI (e.g., I620, I609, I619, S060, S065), while BF FT-g and WHO-g ranked highest the FTs and the disease groups that are involved in cancer (e.g., hsa04068, hsa05217, hsa05205, hsa04350), autoimmune diseases (e.g., hsa05320, dsL43), or indicate central nervous system disorders (e.g, S10, G90, hsa04360, hsa05214). Concerning cardiovascular-relevant predictors, BF WHO-g ranked highest: essential (primary) hypertension (I10) and other disorders of white blood cells (D72). Finally, BF FT-g ranked the target-less predictors group (grp262) highest.

We explored the overlap in selected predictor variables and groups across the methods (see SI_File5). For ungrouped methods, in both samples, LASSO selected the predictors that RF ranked highest. Concerning FT-g, all top 10 FT groups by BF FT-g were also ranked the highest by ARTP FT-g in both samples (except for grp262 in

ICB). NGL FT-g selected FTs and those ranked highest by BF and the ARTP overlapped. Finally, concerning WHO-g, the top 10 WHO groups by BF WHO-g minimally overlapped with the highest ranked groups by ARTP WHO-g or those selected by NGL WHO-g in both GIB and ICB samples.

Moreover, we explored further predictor variables and groups focusing on: 1) adjustment variables (i.e., age, sex and time-to-event), 2) custom-made groups (adjustment variable group grp261 and target-less predictors group grp262), and 3) the DOACs either as predictor drugs or their FT or WHO groups. In GIB sample, age and time-to-event were selected by LASSO, NGL FT-g (split) and NGL WHO-g, while sex was only selected by NGL FT-g (split). Age and time-to-event were ranked among the top 10 predictors by RF but not the ARTP, however the adjustment group of variables (grp261) was ranked high by ARTP FT-g (1gp). In ICB sample showed a similar trend with sex was only selected by NGL FT-g (1gp) (see SI_File5). In GIB sample, aside from BF FT-g results, the three known FT pathways of DOACs, namely, hsa04610 hsa04080 and hsa04810, were ranked the highest only by the ARTP FT-g (1gp and split), while only two of the pathways were selected by NGL FT-g (1gp), also the ARTP WHO-g ranked high the drug group antithrombotic agents (B01A), while the DOAC dabigatran was only selected by the LASSO. In the ICB sample, similar results were obtained, except that no DOACs were selected or ranked high by any method.

4 Discussion

In this study, we compared five statistical methods in their ability to predict an event, comparing non-group-based and group-based methods. We used two grouping schemes, classical WHO classification of drugs and diseases, and pathway-level grouping based on functional target data curated from the TTD. For that, we designed a nested case-control study to construct and analyze two matched case-control subcohorts (1:4) of adult insurants in GePaRD with and without main hospital diagnosis of one of two events, GIB and ICB.

4.1 Events of interest

We chose two serious events that could lead to morbidity and mortality in patients. Here we discuss those events focusing on three aspects: underlying conditions, drug pharmacodynamics and drug pharmacokinetics. Gastrointestinal bleeding is a serious concern in elderly patients. The common causes of GIB in the elderly are underlying conditions (e.g., peptic ulcer, malignancy, diverticular hemorrhage, hemorrhoids, inflammatory bowel disease) [35]. Moreover, both upper and lower GIB is known to be linked to administration of aspirin, nonsteroidal anti-inflammatory drugs (NSAIDs) and antithrombotic drugs (reviewed in[35]). It is, therefore, recommended, in case of GIB, to consider patient history with respect to diseases/procedures (i.e., previous abdominal surgery) and current medication. Concerning ICB, the use of DOACs have been also considered a risk factor for developing ICB in case of mild traumatic brain injury [36]}, increasing in the haemorrhagic risk profile patients under anticoagulant therapy.

In addition to intentional drug targets, the molecules drugs bind to in order to exert their intended effect (see Introduction section), there are the molecules that are involved in drug pharmacokinetics (i.e., absorption, distribution, metabolism, and excretion). We focus on three factors causing pharmacokinetic variations: single nucleotide variants in genes coding for key enzymes in drug pharmacokinetics, co-administration of drugs

interacting with those key enzymes, and pre-existing disease conditions that affect drug absorption and/or elimination [17]. For example. co-administration of the DOACs and drugs that interact with the enzymes integral to DOACs pharmacokinetics can slow down DOACs metabolism and increase risk of bleeding. Those can be either enzymes for activation (CES1 for dabigatran), metabolism (CYP3A5 for apixaban), or clearance (e.g., acetaminophen or morphine interaction with UGT2B15 slow down dabigatran's elimination), or transport proteins [e.g., co-administration of glycoprotein-P inhibitors (e.g., erythromycin, atorvastatin) increase dabigatran's blood concentration]. Single-nucleotide variants in the genes encoding for the aforementioned key enzymes and transporters can lead to variations in DOACs pharmacokinetics and to an increased risk of bleeding [17]. We refer to the comprehensive review by Ašić *et al*. on the pharmacogenetics of DOACs [37].

This study aims at comparing the predictability of an EI in patients given his/her drug exposures and diseases using FTs rather than an in-depth investigation of factors associated with GIB or ICB, or DOACs targets. Below, we focus on discussing the predictive performance of the methods. Nevertheless, we expect to see three categories of predictors (or groups) selected with the highest ranks: event ICDs (e.g., patient history, outpatient diagnosis), ICDs of underlying conditions of the event or those affecting DOACs clearance, and ATCs of the interacting drugs, which also we discuss below.

4.2 Grouping annotation and structure

We inspected several, largely manually curated, biological and chemical databases that link drugs and diseases to FTs, e.g., the comprehensive knowledge base KEGG, TTD focusing on curated drug-target information, and STITCH forg drug-target binding scores. For this model proposed here, we favored the manually curated TTD. Our approach can be also complemented with STITCH data. Issues with FT annotation are mainly: handling predictors of unknown targets and the impact of overlap between groups. The ARTP, FT and NGL analyzed group data independently and were not affected by overlap. It is interesting to evaluate the effect of group size on the selection. Here, the largest groups were not selected (e.g., cancer hsa04020 and metabolic hsa01100 pathways).

Compared to FT-based grouping, conventional ATC/ICD grouping, does not provide the molecular-based interpretation we proposed. However, the group numbers are larger, yet sizes are 2-fold smaller and there is no overlap among the groups. Computationally, this was less challenging for all grouping methods.

4.3 Prediction evaluation

We considered the precision-recall (PR) curve for prediction evaluation as it is argued to be more informative than the receiver operating characteristics (ROC) curve in case of evaluating binary classifiers on imbalanced datasets {Saito2015}. In such EI, we expected highly imbalanced data as only a minor proportion of the predictors are believed to be truely associated with the EI. Based on the PR-AUC, LASSO, RF and BF FT-g performed best in both EI samples. BF is designed for analyzing large groups of omics data. FT-grouping resemble that of variant data, FT-grouping improved BF performance, possibly through increasing the power for detecting association.

WHO-grouping, on the other hand, improved NGL performance. The constructed group variable was based the sum of predictors and correcting for group size. It is possible that, in case of FT-grouping, this resulted in identical

group variables across overlapping groups, which were not handled well by the LASSO. In case of WHO-grouping, there are no overlapping groups and the group variables are somehow unique and would improve the predictive power of the LASSO.

Investigating the performance of a genetic epidemiology method was inspired by the resemblance of our proposed approach to enrichment analysis. The ARTP algorithm is designed for SNP data analysis, in which the number of groups and variables are similar to those we had here, yet the number of observations is expected to be many folds less. The bottleneck in the ARTP was modeling each group, particularly those of very large number of predictors. It is possible that, if computational constraints are addressed, an increase in the number of permutations would improve the ARTP performance.

Finally, the selected predictors by the methods include groups and variables that are known to be linked to the events in question. In case of GIB, concerning WHO-g and ng, RF and BF rank the underlying conditions highest, yet not the drugs. while the LASSO selected a DOAC, and the ARTP ranked antithrombotic agents high. Concerning FT-g, a large overlap between ARTP and BF ranking is observed. It is then important to consider a ranking for the large number of selected predictors (by the LASSO) or groups (by the ARTP), where, respectively, either no *p* value or the same *p* value is generated.

5 Conclusion and Outlook

This study attempted at evaluating statistical methods' performance in predicting ADR risk in health care claims data incorporating molecular ontologies and domain knowledge, and modifying methods transferred from genetic epidemiology for ADR prediction. FT-based grouping would offer an advantage for ADR risk prediction and inference of involved factors, compared to conventional ATC/ICD systems alone, exploiting the underlying relationship between the predictors and the ADR. The results of our comparative study suggest block forests using FTs as a candidate method for individual risk prediction and for inference of suspected risk factors. This study highlights the need for an ad-hoc linear model for quantifying LASSO-generated associations. As well, it highlights considering a cutoff for RF and BF importance, and also considering evaluating both BF blocks based on split value and variables based on importance.

Further investigation is required to determine the extent to which data set size, group structure (i.e., group overlap, handling target-less predictors) and group size affect methods performance. Optimization for methods that correct for group size, such as group LASSO, is also required for large-scale prediction and inference. Moreover, construction of a risk profile as well as using an ensemble risk prediction compiled of more than one method might allow for combining the strengths of those methods, better prediction of ADRs and consequently personalized medical decisions. Our proposed approach for FT-based grouping of predictors can be complemented with drug-drug and, when available, disease-disease relationships as in score matrices (e.g., integrating drug-target score data from STITCH). Furthermore, our model can also be extended with dosage information (i.e., defined daily dose).

Acknowledgements

The authors thank Prof. Dr. Vanessa Didelez, Dr. Tania Schink and Dr. Marvin Wright for their input and fruitful discussions. The authors also thank the statutory health insurance provider which provided data for this study, namely Die Techniker (TK). Finally, the authors gratefully acknowledge the financial support from the innovation fund ("Innovationsfonds") of the Federal Joint Committee in Germany (grant number: 01VSF16020).

References

[1] European Commission. Proposal for a Regulation Amending, As Regards Pharmacovigilance of Medicinal Products for Human Use. 2008; Dec. Regulation (EC) No.: 726/2004.

[2] Meyer UA. Pharmacogenetics and adverse drug reactions. Lancet. 2000;356(9242):1667–71.

[3] Phillips KA, Veenstra DL, Oren E, Lee JK, Sadee W. Potential role of pharmacogenomics in reducing adverse drug reactions: A systematic review. Journal of the American Medical Association. 2001;286(18):2270–2279.

[4] 1. Rizkallah MR, Saad R, Aziz RK. The human microbiome project, personalized medicine and the birth of pharmacomicrobiomics. Current Pharmacogenomics and Personalized Medicine. 2010;8:182–93.

[5] Alomar MJ. Factors affecting the development of adverse drug reactions (Review article). Saudi Pharmaceutical Journal. 2014;22(2):83–94.

[6] Dumbreck S, Flynn A, Nairn M, Wilson M, Treweek S, Mercer SW, et al. Drug-disease and drug-drug interactions: Systematic examination of recommendations in 12 UK national clinical guidelines. BMJ. 2015;350(h949):1–8.

[7] Stewart D, Gibson-Smith K, MacLure K, Mair A, Alonso A, Codina C, et al. A modified Delphi study to determine the level of consensus across the European Union on the structures, processes and desired outcomes of the management of polypharmacy in older people. PLoS ONE. 2017;12(11):1–17.

[8] Schöttker B, Saum KU, Muhlack DC, Hoppe LK, Holleczek B, Brenner H. Polypharmacy and mortality: New insights from a large cohort of older adults by detection of effect modification by multi-morbidity and comprehensive correction of confounding by indication. European Journal of Clinical Pharmacology. 2017;73(8):1041–1048.

[9] Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. Journal of Clinical Epidemiology. 2005;58(4):323–337.

[10] Suling M, Pigeot I. Signal detection and monitoring based on longitudinal healthcare data. Pharmaceutics. 2012;4(4):607–640.

[11] Pacurariu A, Plueschke K, McGettigan P, Morales DR, Slattery J, Vogl D, et al. Electronic healthcare databases in Europe: Descriptive analysis of characteristics and potential for use in medicines regulation. BMJ Open. 2018;8(9):e023090.

[12] Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nature Reviews Drug Discovery. 2006;5(12):993–996.

[13] Yang H, Qin C, Li YH, Tao L, Zhou J, Yu CY, et al. Therapeutic target database update 2016: Enriched resource for bench to clinical drug target and targeted pathway information. Nucleic Acids Research. 2016;44(D1):D1069–D1074.

[14] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: A large-scale bioactivity database for drug discovery. Nucleic Acids Research. 2012;40(D1):1100–1107.

[15] Szklarczyk D, Santos A, Von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Research. 2016;44(D1):D380–D384.

[16] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research. 2017;45(D1):D353–D361.

[17] Shnayder NA, Petrova MM, Shesternya PA, Savinova AV, Bochanova EN, Zimnitskaya OV, et al. Using pharmacogenetics of direct oral anticoagulants to predict changes in their pharmacokinetics and the risk of adverse drug reactions. Biomedicines. 2021;9(5):451.

[18] Vilar S, Harpaz R, Santana L, Uriarte E, Friedman C. Enhancing adverse drug event detection in electronic health records using molecular structure similarity: Application to pancreatitis. PLoS ONE. 2012;7(7):e41471.

[19] Liu M, Wu Y, Chen Y, Sun J, Zhao Z, Chen Xw, et al. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. Journal of the American Medical Informatics Association. 2012;19(e1):e28–e35.

[20] Liu R, AbdulHameed MDM, Kumar K, Yu X, Wallqvist A, Reifman J. Data-driven prediction of adverse drug reactions induced by drug-drug interactions. BMC Pharmacology and Toxicology. 2017;18(1):1–18.

[21] Saunders G, Ivkovic S, Ghosh R, Yearwood J. Applying anatomical therapeutic chemical (ATC) and critical term ontologies to Australian drug safety data for association rules and adverse event signalling. In: Proceedings of the 2005 Australasian Ontology Workshop - Volume 58. AOW '05. Darlinghurst, Australia: Australian Computer Society, Inc.; 2005. p. 93–98.

[22] Winnenburg R, Sorbello A, Bodenreider O. Exploring adverse drug events at the class level. Journal of Biomedical Semantics. 2015;6(18):1–10.

[23] Arnaud M, Bégaud B, Thurin N, Moore N, Pariente A, Salvo F. Methods for safety signal detection in healthcare databases: A literature review. Expert Opinion on Drug Safety. 2017;16(6):721–732.

[24] Pigeot I, Ahrens W. Establishment of a pharmacoepidemiological database in Germany: Methodological potential, scientific value and practical limitations.Pharmacoepidemiology and Drug Safety. 2008;17(3):215–223.

[25] Li YH, Yu CY, Li XX, Zhang P, Tang J, Yang Q, et al. Therapeutic target database update 2018: Enriched resource for facilitating bench-to-clinic research of targeted therapeutics. Nucleic Acids Research. 2018;46(D1):D1121–D1127.

[26] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;58(1):267–288.

[27] Caster O, Norén GN, Madigan D, Bate A. Large-scale regression-based pattern discovery in international adverse drug reaction surveillance. In: Proceedings of the KDD-08 Workshop on Mining MedicalData; 2008. p. S 24–27.

[28] Breiman L. Random forests. Machine Learning. 2001;45(1):5–32.

[29] Hornung R, Wright MN. Block forests: Random forests for blocks of clinical and omics covariate data. BMC Bioinformatics. 2019;20(1):1–17.

[30] Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, et al. Pathway analysis by adaptive combination of p-values. Genetic Epidemiology. 2009;33(8):700–709.

[31] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010;33(1):1–22.

[32] Wright MN, Ziegler A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software. 2017;77(1):1–17.

[33] Dijkstra L, Rizkallah MR, Foraita R. ARTPredict: Binary outcome prediction using the adaptive combination of p-values [Internet]. GitHub; 2021 [cited 2021 Nov 13]. Available from: https://github.com/bips-hb/ARTPredict

[34] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3):1–21.

[35] Yachimski PS, Friedman LS. Gastrointestinal bleeding in the elderly. Nature Clinical Practice Gastroenterology and Hepatology. 2008;5(2):80–93.

[36] Turcato G, Zaboli A, Zannoni M, Ricci G, Zorzi E, Ciccariello L, et al. Risk factors associated with intracranial bleeding and neurosurgery in patients with mild traumatic brain injury who are receiving direct oral anticoagulants. American Journal of Emergency Medicine. 2021;43:180–185.

[37] Ašić A, Marjanović D, Mirat J, Primorac D. Pharmacogenetics of novel oral anticoagulants: A review of identified gene variants & future perspectives. Personalized Medicine. 2018;15(3):209–221.

Tables

Table 1: Descriptive statistics of the study populations.

	GIB			ICB		
	Affected cases (N=12944)	**Controls (N=51776)**	**Total (N=64720)**	**Affected cases (N=6920)**	**Controls (N=27680)**	**Total (N=34600)**
Sex						
Male (%)	7767 (60%)	31068 (60%)	38835 (60%)	4319 (62.4%)	17276 (62.4%)	21595 (62.4%)
Female (%)	5177 (40%)	20708 (40%)	25885 (40%)	2601 (37.6%)	10404 (37.6%)	13005 (37.6%)
Age, yrs (Range 18 - 101)						
Mean (SD)	66 (17)	66 (17)	66 (17)	67 (16)	67 (16)	67 (16)
Time-to-event, days (Range 90 - 730)						
Mean (SD)			417 (185)			411 (185)
Death						
Yes (%)	352 (2.7%)	0 (0%)	352 (0.5%)	560 (8.1%)	0 (0%)	560 (1.6%)
No. covariates						
Total	8040	7346	8577	6303	7338	7847
Mean (SD)	48.2 (29.6)	27.1 (22.5)	31.3 (25.5)	42.3 (27)	27.1 (22.7)	30.1 (24.4)
Range	0 - 310	0 - 221	0 - 310	0 - 218	0 - 222	0 - 222
No. NI (%)	537 (6.3)	1231 (14.4)	0 (0)	1544 (19.7)	509 (6.5)	0 (0)
DOACs use						
Yes (%)	1596 (12.3)	2342 (4.5)	3938 (6.1)	485 (7)	1253 (4.5)	1738 (5)

GIB = Gastrointestinal bleeding; ICB = intracranial bleeding; NI = Non-informative covariates; DOACs = Direct oral anticoagulants.

Table 2: The area under the precision-recall curve (PR-AUC) of prediction performance for GIB and ICB.

	GIB	ICB
Ungrouped		
LASSO	0.706	0.798
RF	0.703	0.806
Grouped (FT-g)		
NGL (as 1gp; split)	0.582; 0.556	0.667; 0.617
BF (as 1gp; split)	0.702; NA	0.825; NA
ARTP (as 1gp; split)	0.449; 0.446	0.312; 0.311
Grouped (WHO-g)		
NGL (as 1gp; split)	0.683; NA	0.8; NA
BF (as 1gp; split)	0.53; NA	0.579; NA
ARTP (as 1gp; split)	0.515; NA	0.596; NA
Ensemble		
ng + FT-g (as 1gp)	0.716	0.824
ng + WHO-g	0.712	0.82
ng + FT-g (1gp) + WHO-g	0.713	0.821

GIB = Gastrointestinal bleeding; ICB = intracranial bleeding; RF = Random Forest; NGL = Naïve-group LASSO; BF = Block Forest; ARTP = Adaptive combination of Rank Truncated Product; FT-g = Functional Target-based grouping; WHO-g = ATC/ICD-based grouping; ng = ungrouped; 1gp = Target-less singleton predictors grouped as one group; split = Target-less singleton predictors treated split into single groups of one predictor each. PR-AUC baseline = 0.2 (see Methods).

Table 3: Number of selected predictors for GIB and ICB.

	GIB	ICB
Ungrouped		
LASSO	840	107
RF[1]	100	100
Grouped (FT-g)		
NGL (as 1gp; split)	187; 978	192; 52
BF[1] *(as 1gp; split)*	100; NA	100; NA
ARTP (as 1gp; split)	256; 687	254; 482
Grouped (WHO-g)		
NGL (as 1gp; split)	487; NA	359; NA
BF[1] *(as 1gp; split)*	100; NA	100; NA
ARTP (as 1gp; split)	871; NA	523; NA
Ensemble		
ng + FT-g (as 1gp)	1158	437
ng + WHO-g	1965	899
ng + FT-g (1gp) + WHO-g	2259	1187

[1]only top 100 predictors were considered

GIB = Gastrointestinal bleeding; ICB = intracranial bleeding; RF = Random Forest; NGL = Naïve-group LASSO; BF = Block Forest; ARTP = Adaptive combination of Rank Truncated Product; FT-g = Functional Target-based grouping; WHO-g = ATC/ICD-based grouping; ng = ungrouped; 1gp = Target-less singleton predictors grouped as one group; split = Target-less singleton predictors treated split into single groups of one predictor each.

Table 4: Highest ranking predictors and groups, according to importance (RF) or split value (BF) for GIB and ICB. For RF, variable refers to ICD code. For BF (FT-g 1gp), group refers to KEGG pathway ID. For BF (WHO-g 1gp), group refers to ICD group (prefix: ds).

4a: GIB

RF		BF (FT-g 1gp)		BF (WHO-g 1gp)	
Variable	**Description**	**Group**	**Description**	**Group**	**Description**
D62	Acute posthemorrhagic anemia	hsa05219	Bladder cancer	dsR43	Disturbances of smell and taste
K922	Gastrointestinal hemorrhage, unspecified	hsa04115	p53 signaling pathway	dsD41	Neoplasm of uncertain behavior of urinary organs
K921	Melena	hsa05120	Epithelial cell signaling in Helicobacter pylori infection	dsQ91	Trisomy 18 and Trisomy 13
K920	Hematemesis	mtu03020	RNA polymerase	dsH92	Otalgia and effusion of ear
K298	Duodenitis	hsa04920	Adipocytokine signaling pathway	dsI43	Cardiomyopathy in diseases classified elsewhere
K290	Acute gastritis	hsa04730	Long-term depression	dsD89	Oth disorders involving the immune mechanism, NEC
D500	Iron deficiency anemia secondary to blood loss (chronic)	hsa04610	Complement and coagulation cascades	dsD56	Thalassemia
K226	Gastro-esophageal laceration-hemorrhage syndrome	hsa05212	Pancreatic cancer	dsD84	Other immunodeficiencies
K625	Hemorrhage of anus and rectum	hsa04911	Insulin secretion	dsF28	Oth psych disorder not due to a sub or known physiol cond
K296	Other gastritis	hsa04614	Renin-angiotensin system	dsC24	Malignant neoplasm of other and unsp parts of biliary tract

4b: ICB

RF		BF (FT-g 1gp)		BF (WHO-g 1gp)	
Variable	**Description**	**Group**	**Description**	**Group**	**Description**
G810	Flaccid hemiplegia	hsa04068	FoxO signaling pathway	dsH74	Other disorders of middle ear mastoid
S065	Traumatic subdural hemorrhage	hsa00410	beta-Alanine metabolism	dsI10	Essential (primary) hypertension
I620	Nontraumatic subdural hemorrhage	hsa05217	Basal cell carcinoma	dsD72	Other disorders of white blood cells
S066	Traumatic subarachnoid hemorrhage	hsa04360	Axon guidance	dsL43	Lichen planus
G936	Cerebral edema	hsa05214	Glioma	dsS10	Superficial injury of neck
I609	Nontraumatic subarachnoid hemorrhage, unspecified	hsa04350	TGF-beta signaling pathway	dsY36	Operations of war
S020	Fracture of vault of skull	hsa05320	Autoimmune thyroid disease	dsG90	Disorders of autonomic nervous system
I619	Nontraumatic intracerebral hemorrhage, unspecified	grp262	"Singletons group"	dsZ48	Encounter for other postprocedural aftercare
R412	Retrograde amnesia	hsa00592	alpha-Linolenic acid metabolism	dsO41	Other disorders of amniotic fluid and membranes
S060	Concussion	hsa05205	Proteoglycans in cancer	dsI62	Other and unspecified nontraumatic intracranial hemorrhage

GIB = Gastrointestinal bleeding; ICB = intracranial bleeding; RF = Random Forest; BF = Block Forest; FT-g = Functional Target-based grouping; WHO-g = ATC/ICD-based grouping;1gp = Target-less singleton predictors grouped as one group.

Figure legends

Figure 1. A schematic representation of two approaches for ADR risk prediction. (a) illustrates the standard approach in the field of pharmacovigilance. The left column contains all the drugs *(1, 2, . . . , k)* and diseases *(1, 2, . . . , l)*. The ADR of interest is shown on the right. The predictions are based on the associations between individual risk factors (drugs and diseases) and the ADR, represented here by arrows pointing from each drug/disease to the ADR. (b) illustrates the approach proposed in this study. Similarly, the left column contains all the drugs *(1, 2, . . . , k)* and diseases *(1, 2, . . . , l)* as covariates. The middle column lists groups *(1, 2, . . . , G)* . Each arrow between a drug/disease and a group represents the group membership. Note that drugs/diseases can belong to multiple groups simultaneously, e.g., $drug_2$ is in, both, $group_1$ and $group_2$. Instead of assessing the associations between the drugs/diseases and the ADR directly as in (a), the associations between the groups and the ADR are assessed, shown here by arrows pointing from the groups to the ADR.

Figure 2. The proposed approach to predict ADRs in routine data of the SHIs using functional targets (FTs). First, relevant online genomic knowledge bases are queried for drug-target, disease-target, drug-disease and drug-drug relationships to curate FTs. FTs serve as the grouping structure of the predictors. Within those FTs, substructures and pairings exist, such as drug-drug structural and functional similarity, drug-disease relationship, and less likely disease-disease co-existence. Second, an epidemiological study design is considered, and SHIs database (here GePaRD) is queried for prescribed drugs, in- and outpatient diagnoses that are coded according to international coding systems to facilitate being mapped to FTs. Third, the SHIs predictors are grouped according to the grouping structure, and the risk of ADR is predicted using statistical models based on those structures. Drugs are denoted in blue, while diseases are in green. Within a FT, solid lines represent drug-drug or disease-disease relationships; dotted lines represent drug-disease (i.e., indication) relationships. GePaRD = The German Pharmacoepidemiological Research Database; ADR = Adverse Drug Reaction.

Figure 3. Study eligibility and matching flowchart. The flowchart illustrates the number of available insurants in the presented nested case-control cohort study, and the number of cases and controls in each subcohort. GIB = Gastrointestinal bleeding; ICB = Intracranial bleeding.

Figure 4. Frequencies of group sizes in study samples. The distribution of the number of non-zero variance drugs and diseases predictors per group in each of the two grouping schemes for each of the study samples: gastrointestinal bleeding (GIB) and intracranial bleeding (ICB). Top-left: Frequencies of predictors per group size category grouped according to functional targets from the Therapeutic Target Database (TTD) in GIB sample; top-right: Frequencies of predictors per group size category grouped according to functional targets from the TTD in ICB sample; bottom-left: Frequencies of predictors per group size category grouped according to the WHO grouping of drugs and diseases as in the ATC/ICD classification in GIB sample; bottm-right = Frequencies of predictors per group size category grouped according to the WHO grouping of drugs and diseases as in the ATC/ICD classification in ICB sample.

Supplementary information legends

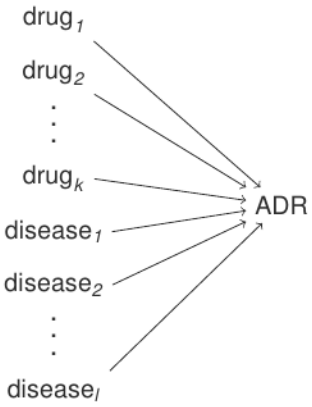
SI_File1: Table of ICD codes considered as outcome in the study, their frequency and explanation.

SI_File2: Table of detailed descriptive statistics of the study populations.

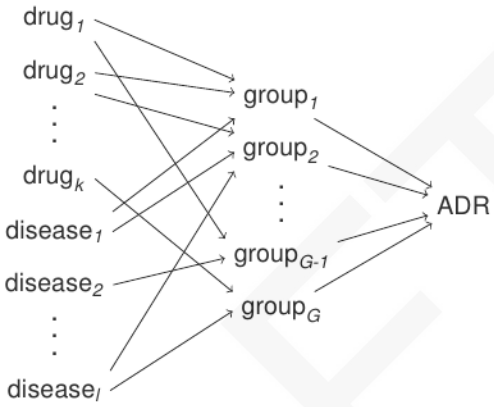
SI_File3: Table of model performance for gastrointestinal bleeding and intracranial bleeding samples. Singleton predictors that did not belong to a functional target (FT) were either analyzed as one group (one gp), or split into dummy groups of one (split gps).

SI_File4: Information on variables and groups selected for gastrointestinal bleeding (GIB) and intracranial bleeding (ICB) subcohorts by random forests (RF) and block forests (BF). Table 1 contains the highest ranked top 10 variables or groups, description and, respectively, importance value or split value. Table 2 shows the overlap of the variables and groups selected by all the methods and settings for GIB. Table 3 show the overlap of the variables and groups selected by all the methods and settings for ICB.

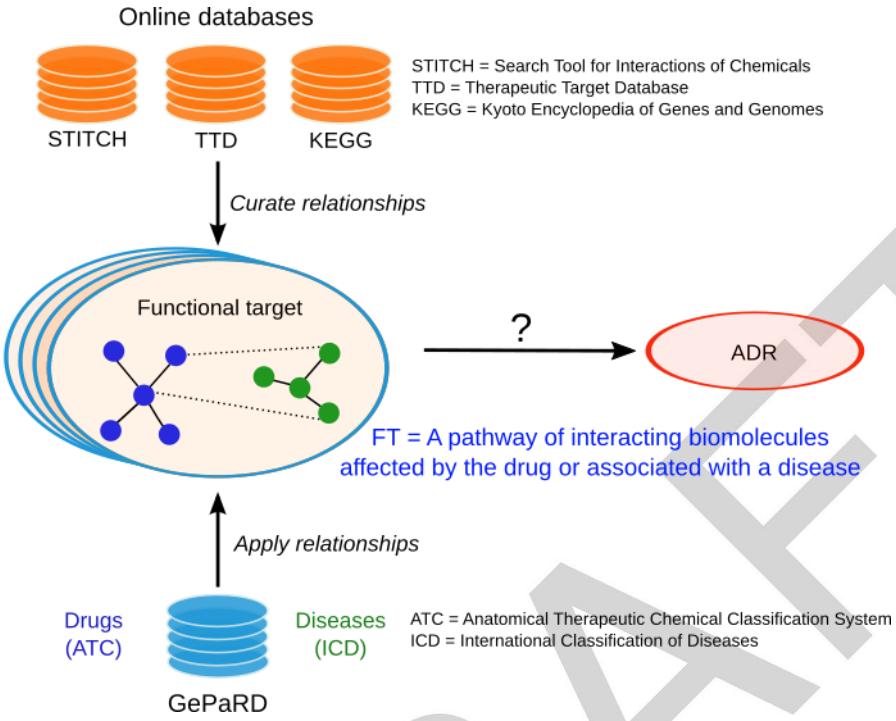
Figures

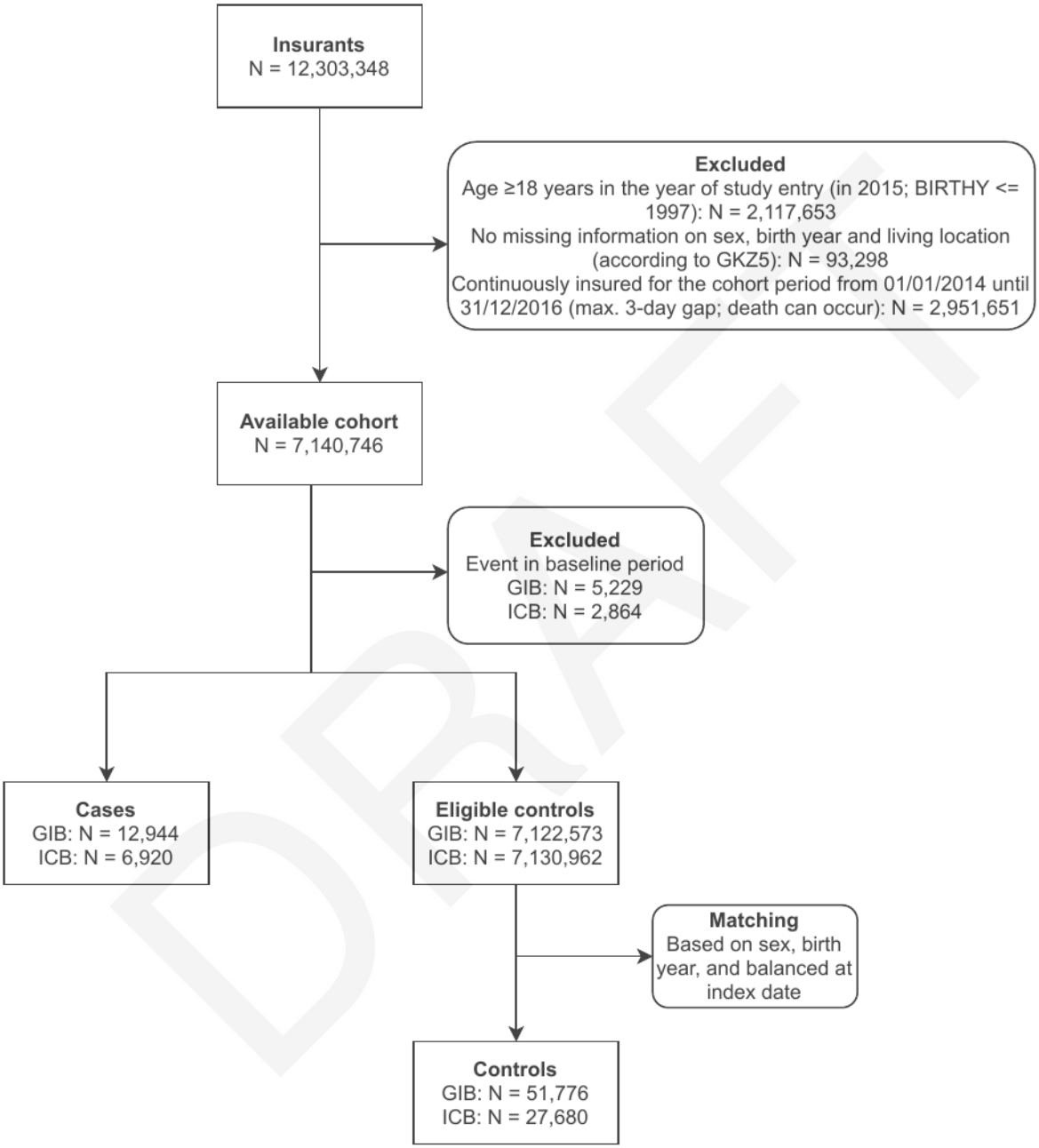


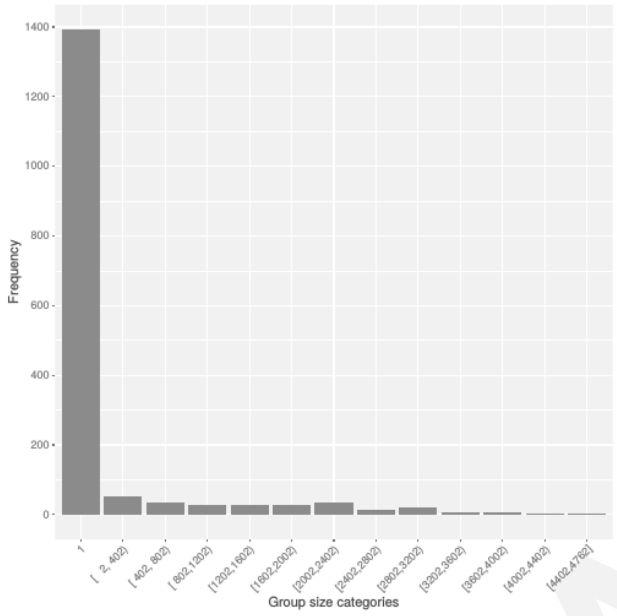
(a) without groups

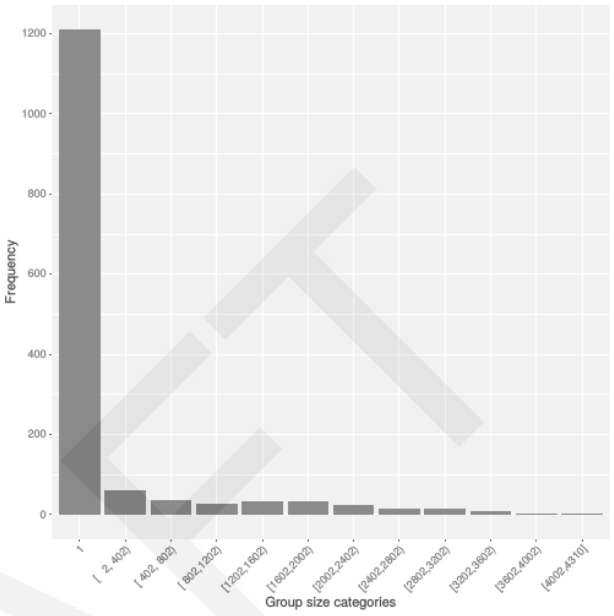


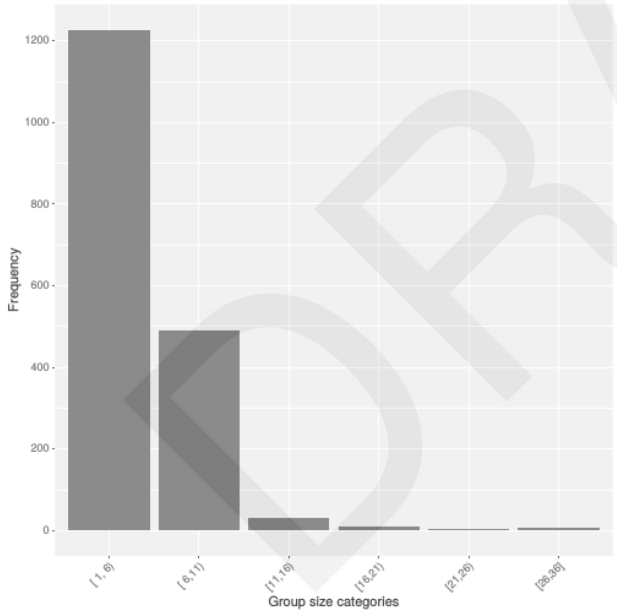
(b) with groups

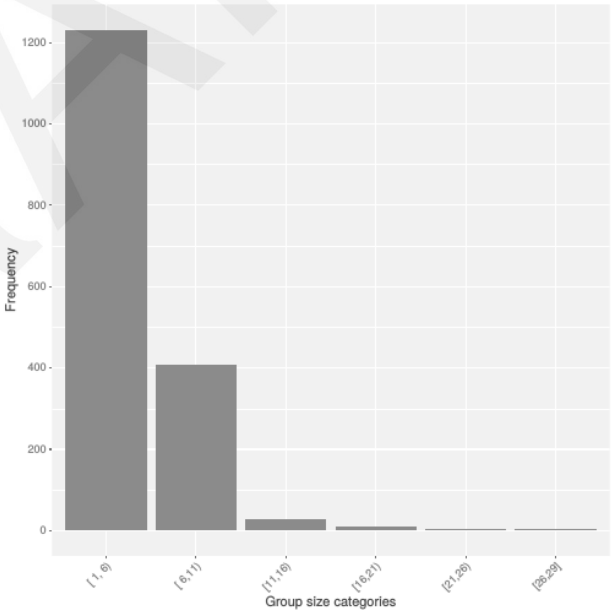

Online databases
STITCH
TTD
KEGG
STITCH = Search Tool for Interactions of Chemicals
TTD = Therapeutic Target Database
KEGG = Kyoto Encyclopedia of Genes and Genomes
Curate relationships
Functional target
?
ADR
FT = A pathway of interacting biomolecules
affected by the drug or associated with a disease
Apply relationships
Drugs
(ATC)
GePaRD
Diseases
(ICD)
ATC = Anatomical Therapeutic Chemical Classification System
ICD = International Classification of Diseases


Insurants
N = 12,303,348
Excluded
Age ≥18 years in the year of study entry (in 2015; BIRTHY <= 1997): N = 2,117,653
No missing information on sex, birth year and living location (according to GKZ5): N = 93,298
Continuously insured for the cohort period from 01/01/2014 until 31/12/2016 (max. 3-day gap; death can occur): N = 2,951,651
Available cohort
N = 7,140,746
Excluded
Event in baseline period
GIB: N = 5,229
ICB: N = 2,864
Cases
GIB: N = 12,944
ICB: N = 6,920
Eligible controls
GIB: N = 7,122,573
ICB: N = 7,130,962
Matching
Based on sex, birth year, and balanced at index date
Controls
GIB: N = 51,776
ICB: N = 27,680


Frequency
1400
1200
1000
800
600
400
200
0
1
[2, 402)
[402, 802)
[802,1202)
[1202,1602)
[1602,2002)
[2002,2402)
[2402,2802)
[2802,3202)
[3202,3602)
[3602,4002)
[4002,4402)
[4402,4782]
Group size categories


Frequency
1200
1000
800
600
400
200
0
1
[2, 402)
[402, 802)
[802,1202)
[1202,1602)
[1602,2002)
[2002,2402)
[2402,2802)
[2802,3202)
[3202,3602)
[3602,4002)
[4002,4310]
Group size categories


Frequency
1200
1000
800
600
400
200
0
[1, 6)
[6,11)
[11,16)
[16,21)
[21,26)
[26,36]
Group size categories


Frequency
1200
1000
800
600
400
200
0
[1, 6)
[6,11)
[11,16)
[16,21)
[21,26)
[26,29]
Group size categories

Bibliography

Acharjee A (2012). Comparison of regularized regression methods for ˜omics data. *Journal of Postgenomics Drug & Biomarker Development* **03**(03): 126.

Ahrens W, Siani A, Adan R, De Henauw S, Eiben G, Gwozdz W, Hebestreit A, Hunsberger M, Kaprio J, Krogh V, Lissner L, Molnár D, Moreno LA, Page A, Picó C, Reisch L, Smith RM, Tornaritis M, Veidebaum T, Williams G, Pohlabeln H, Pigeot I & on behalf of the I.Family consortium (2017). Cohort profile: The transition from childhood to adolescence in European children–how I.Family extends the IDEFICS cohort. *International Journal of Epidemiology* **46**(5): 1394–1395j.

Alderkamp AC, Kulk G, Buma AGJ, Visser RJW, Van Dijken GL, Mills MM & Arrigo KR (2012). The effect of iron limitation on the photophysiology of *Phaeocystis antarctica* (Prymnesiophyceae) and *Fragilariopsis cylindrus* (Bacillariophyceae) under dynamic irradiance. *Journal of Phycology* **48**(1): 45–59.

Allen AE, Laroche J, Maheswari U, Lommer M, Schauer N, Lopez PJ, Finazzi G, Fernie AR & Bowler C (2008). Whole-cell response of the pennate diatom *Phaeodactylum tricornutum* to iron starvation. *Proceedings of the National Academy of Sciences* **105**(30): 10438–10443.

Alomar MJ (2014). Factors affecting the development of adverse drug reactions. *Saudi Pharmaceutical Journal* **22**(2): 83–94.

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**(3): 403–410.

Andrews EB, Margulis AV, Tennis P & West SL (2014). Opportunities and challenges in using epidemiologic methods to monitor drug safety in the era of large automated health databases. *Current Epidemiology Reports* **1**(4): 194–205.

Angiuoli SV, Gussman A, Klimke W, Cochrane G, Field D, Garrity GM, Kodira CD, Kyrpides N, Madupu R, Markowitz V, Tatusova T, Thomson N & White O (2008). Toward an online repository of standard operating procedures (SOPs) for (meta)genomic annotation. *OMICS: A Journal of Integrative Biology* **12**(2): 137–141.

Arnaud M, Bégaud B, Thurin N, Moore N, Pariente A & Salvo F (2017). Methods for safety signal detection in healthcare databases: A literature review. *Expert Opinion on Drug Safety* **16**(6): 721–732.

Assmy P, Henjes J, Klaas C & Smetacek V (2007). Mechanisms determining species dominance in a phytoplankton bloom induced by the iron fertilization experiment

EisenEx in the Southern Ocean. *Deep Sea Research Part I: Oceanographic Research Papers* **54**(3): 340–362.

Baker SB, Xiang W & Atkinson I (2017). Internet of Things for smart healthcare: technologies, challenges, and opportunities. *IEEE Access* **5**: 26521–26544.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S & Soboleva A (2013). NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Research* **41**(D1): 991–995.

Bayraktarov E, Ehmke G, O'Connor J, Burns EL, Nguyen HA, McRae L, Possingham HP & Lindenmayer DB (2019). Do big unstructured biodiversity data mean more knowledge? *Frontiers in Ecology and Evolution* **6**: 239.

Behnke J & LaRoche J (2020). Iron uptake proteins in algae and the role of iron starvation-induced proteins (ISIPs). *European Journal of Phycology* **55**(3): 339–360.

Bender SJ, Moran DM, McIlvin MR, Zheng H, McCrow JP, Badger J, DiTullio GR, Allen AE & Saito MA (2018). Iron triggers colony formation in *Phaeocystis antarctica*: Connecting molecular mechanisms with iron biogeochemistry. *Biogeosciences* **15**(16): 4923–4942.

Bengtsson H (2019). *future: Unified parallel and distributed processing in R for everyone.* R package version 1.14.0. URL: https://CRAN.R-project.org/package=future.

Bhattacharya I & Getoor L (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data* **1**(1): 5–es.

Biswal B, Joshi PN, Raval MK & Biswal UC (2011). Photosynthesis, a global sensor of environmental stress in green plants: Stress signalling and adaptation. *Current Science* **101**(1): 47–56.

Bizer C, Boncz P, Brodie ML & Erling O (2012). The meaningful use of big data. *ACM SIGMOD Record* **40**(4): 56–60.

Bolger AM, Lohse M & Usadel B (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15): 2114–2120.

Börnhorst C, Russo P, Veidebaum T, Tornaritis M, Molnár D, Lissner L, Marild S, De Henauw S, Moreno LA, Intemann T, Wolters M, Ahrens W & Floegel A (2019). Metabolic status in children and its transitions during childhood and adolescence—the IDEFICS/I.Family study. *International Journal of Epidemiology* **48**(5): 1673–1683.

Borowitzka MA (2018). The 'stress' concept in microalgal biology—homeostasis, acclimation and adaptation. *Journal of Applied Phycology* **30**(5): 2815–2825.

Boyd PW (2002a). Environmental factors controlling phytoplankton processes in the Southern Ocean. *Journal of Phycology* **38**(5): 844–861.

Boyd PW (2002b). The role of iron in the biogeochemistry of the Southern Ocean and equatorial Pacific: A comparison of *in situ* iron enrichments. *Deep Sea Research Part II: Topical Studies in Oceanography* **49**(9-10): 1803–1821.

Breheny P (2015). The group exponential lasso for bi-level variable selection. *Biometrics* **71**(3): 731–740.

Breheny P & Huang J (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface* **2**(3): 369–380.

Bremer AA, Mietus-Snyder M & Lustig RH (2012). Toward a unifying hypothesis of metabolic syndrome. *Pediatrics* **129**(3): 557–570.

Bressa C, Bailén-Andrino M, Pérez-Santiago J, González-Soltero R, Pérez M, Montalvo-Lominchar MG, Maté-Muñoz JL, Domínguez R, Moreno D & Larrosa M (2017). Differences in gut microbiota profile between women with active lifestyle and sedentary women. *PLoS ONE* **12**(2): 1–20.

Brown ED (2014). Drowning in Data, Starved for Information, *Eric D. Brown, D.Sc.* URL: https://ericbrown.com/drowning-in-data-starved-for-information.htm (Accessed on: 2021-11-08).

Bruce P & Bruce A (2017). *Practical Statistics for Data Scientists* O'Reilly Media, Inc., Sebastopol, CA.

Buneman P, Davidson S, Fernandez M & Suciu D (1996). Adding structure to unstructured data Technical Report No. MS-CIS-96-21, University of Pennsylvania Department of Computer and Information Science, Pennsylvania, US.

Buneman P, Davidson SB & Suciu D (1995). Programming constructs for unstructured data In *Proceedings of the Fifth International Workshop on Database Programming Languages*, DBLP-5. Springer, Berlin, Heidelberg.

Burns P (2011). *The R Inferno* lulu, 1st edition Available from: https://www.burns-stat.com/pages/Tutor/R_inferno.pdf.

Chan EW, Liu KQL, Chui CSL, Sing CW, Wong LYL & Wong ICK (2015). Adverse drug reactions - examples of detection of rare events using databases. *British Journal of Clinical Pharmacology* **80**(4): 855–861.

Chan Y, Talburt J & Talley TM, editors (2010). *Data Engineering: Mining, Information and Intelligence* Vol. 132 of *International Series in Operations Research & Management Science* Springer, Boston, MA.

Charrad M, Ghazzali N, Boiteau V & Niknafs A (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software* **61**(6): 1–36.

Checa A, Bedia C & Jaumot J (2015). Lipidomic data analysis: Tutorial, practical guidelines and applications. *Analytica Chimica Acta* **885**: 1–16.

Cheng JM, Suoniemi M, Kardys I, Vihervaara T, de Boer SPM, Akkerhuis KM, Sysi-Aho M, Ekroos K, Garcia-Garcia HM, Oemrawsingh RM, Regar E, Koenig W, Serruys PW, van Geuns RJ, Boersma E & Laaksonen R (2015). Plasma concentrations of molecular lipid species in relation to coronary plaque characteristics and cardiovascular outcome: Results of the ATHEROREMO-IVUS study. *Atherosclerosis* **243**(2): 560–566.

Chua HH, Chou HC, Tung YL, Chiang BL, Liao CC, Liu HH & Ni YH (2018). Intestinal dysbiosis featuring abundance of *Ruminococcus gnavus* associates with allergic diseases in infants. *Gastroenterology* **154**(1): 154–167.

Coate JE & Doyle JJ (2010). Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: An example from a plant allopolyploid. *Genome Biology and Evolution* **2**(0): 534–546.

Cock PJA, Fields CJ, Goto N, Heuer ML & Rice PM (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* **38**(6): 1767–1771.

Cohen NR, Gong W, Moran DM, McIlvin MR, Saito MA & Marchetti A (2018). Transcriptomic and proteomic responses of the oceanic diatom *Pseudo-nitzschia granii* to iron limitation. *Environmental Microbiology* **20**(8): 3109–3126.

Cole TJ & Lobstein T (2012). Extended international (IOTF) body mass index cut-offs for thinness, overweight and obesity. *Pediatric Obesity* **7**(4): 284–294.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X & Mortazavi A (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology* **17**(1): 1–19.

Coppola L, Cianflone A, Grimaldi AM, Incoronato M, Bevilacqua P, Messina F, Baselice S, Soricelli A, Mirabelli P & Salvatore M (2019). Biobanking in health care: Evolution and future directions. *Journal of Translational Medicine* **17**(1): 172.

Cosgriff CV, Ebner DK & Celi LA (2020). Data sharing in the era of COVID-19. *The Lancet Digital Health* **2**(5): e224.

Cotter D, Maer A, Guda C, Saunders B & Subramaniam S (2006). LMPD: LIPID MAPS proteome database. *Nucleic Acids Research* **34**(Database issue): D507–D510.

Cox DR (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)* **20**(2): 215–232.

Cugino D, Gianfagna F, Ahrens W, De Henauw S, Koni AC, Marild S, Molnar D, Moreno LA, Pitsiladis Y, Russo P, Siani A, Tornaritis M, Veidebaum T & Iacoviello L (2013). Polymorphisms of matrix metalloproteinase gene and adiposity indices in European children: Results of the IDEFICS study. *International Journal of Obesity* **37**(12): 1539–1544.

Curcin V, Ghanem M, Molokhia M, Guo Y & Darlington J (2008). Mining adverse drug reactions with e-Science workflows In *2008 Cairo International Biomedical Engineering Conference*, pp. 1–5. IEEE.

Dai M, Thompson RC, Maher C, Contreras-Galindo R, Kaplan MH, Markovitz DM, Omenn G & Meng F (2010). NGSQC: Cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics* **11**(Suppl 4): S7.

Datta S & Mertens BJA, editors (2017). *Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry* Number February 2019 in Frontiers in Probability and the Statistical Sciences. Springer, Cham.

Davidson NM & Oshlack A (2018). Necklace: Combining reference and assembled transcriptomes for more comprehensive RNA-Seq analysis. *GigaScience* **7**(5): 1–6.

de Baar HJW, Boyd PW, Coale KH, Landry MR, Tsuda A, Assmy P, Bakker DCE, Bozec Y, Barber RT, Brzezinski MA, Buesseler KO, Boyé M, Croot PL, Gervais F, Gorbunov MY, Harrison PJ, Hiscock WT, Laan P, Lancelot C, Law CS, Levasseur M, Marchetti A, Millero FJ, Nishioka J, Nojiri Y, van Oijen T, Riebesell U, Rijkenberg MJA, Saito H, Takeda S, Timmermans KR, Veldhuis MJW, Waite AM & Wong CS (2005). Synthesis of iron fertilization experiments: From the iron age in the age of enlightenment. *Journal of Geophysical Research* **110**(C9): C09S16.

De Mauro A, Greco M & Grimaldi M (2015). What is big data? A consensual definition and a review of key research topics. *AIP Conference Proceedings* **1644**(1): 97–104.

De Mello VD, Lankinen M, Schwab U, Kolehmainen M, Lehto S, Seppänen-Laakso T, Orešič M, Pulkkinen L, Uusitupa M & Erkkilä AT (2009). Link between plasma ceramides, inflammation and insulin resistance: Association with serum IL-6 concentration in patients with coronary heart disease. *Diabetologia* **52**(12): 2612–2615.

Dean J & Ghemawat S (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM* **51**(1): 107–113.

Denny JC & Collins FS (2021). Precision medicine in 2030—seven ways to transform healthcare. *Cell* **184**(6): 1415–1419.

Dey N, Hassanien AE, Bhatt C, Ashour AS & Satapathy SC, editors (2018). *Internet of Things and Big Data Analytics Toward Next-Generation Intelligence* Vol. 30 of *Studies in Big Data* Springer, Cham.

Dimitrov DV (2016). Medical internet of things and big data in healthcare. *Healthcare Informatics Research* **22**(3): 156–163.

Dixon P (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* **14**(6): 927–930.

Dugdale R & Wilkerson F (1991). Low specific nitrate uptake rate: A common feature of high-nutrient, low-chlorophyll marine ecosystems. *Limnology and Oceanography* **36**(8): 1678–1688.

Dumbreck S, Flynn A, Nairn M, Wilson M, Treweek S, Mercer SW, Alderson P, Thompson A, Payne K & Guthrie B (2015). Drug-disease and drug-drug interactions: Systematic examination of recommendations in 12 UK national clinical guidelines. *BMJ* **350**: h949.

Eder K, Baffy N, Falus A & Fulop AK (2009). The major inflammatory mediator interleukin-6 and obesity. *Inflammation Research* **58**(11): 727–736.

El Emam K, Dankar FK, Issa R, Jonker E, Amyot D, Cogo E, Corriveau JP, Walker M, Chowdhury S, Vaillancourt R, Roffey T & Bottomley J (2009). A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association* **16**(5): 670–682.

Enders D (2017). Designs and Analytical Strategies to Control for Unmeasured Confounding in Studies Based on Administrative Health Care Databases Ph.D. diss., Universität Bremen URL: https://elib.suub.uni-bremen.de/peid/D00105995.html.

Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, Goldman N, Hubbard TJ, Harrow J, Guigó R, Bertone P, Alioto T, Behr J, Bohnert R, Campagna D, Davis CA, Dobin A, Gingeras TR, Jean G, Kosarev P, Li S, Liu J, Mason CE, Molodtsov V, Ning Z, Ponstingl H, Prins JF, Ribeca P, Seledtsov I, Solovyev V, Valle G, Vitulo N, Wang K, Wu TD & Zeller G (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods* **10**(12): 1185–1191.

European Commission (2008). Proposal for a regulation amending, as regards pharmacovigilance of medicinal products for human use. Regulation (EC) No 726/2004, Commission of The European Communities, Brussels, BE.

Fanaee-T H & Thoresen M (2019). Performance evaluation of methods for integrative dimension reduction. *Information Sciences* **493**: 105–119.

Ferrari C, Proost S, Ruprecht C & Mutwil M (2018). PhytoNet: Comparative co-expression network analyses across phytoplankton and land plants. *Nucleic Acids Research* **46**(W1): W76–W83.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA & Merrick JM (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**(5223): 496–512.

Fogg GE (2001). *Algal Adaptation to Environmental Stresses*, pp. 1–19 Springer, Berlin, Heidelberg.

for Standardization IO (2019). Information technology database languages — SQL — Part 15: Multi-dimensional arrays (SQL/MDA) Standard ISO/IEC 9075-15:2019, International Organization for Standardization, Geneva, CH.

Foraita R, Dijkstra L, Falkenberg F, Garling M, Linder R, Pflock R, Rizkallah MR, Schwaninger M, Wright MN & Pigeot I (2018). Aufdeckung von Arzneimittelrisiken nach der Zulassung. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz* **61**(9): 1075–1081.

Friedman J, Hastie T & Tibshirani R (2010). A note on the group lasso and a sparse group lasso. arXiv: 1001.0736.

Frolova A & Obolenska M (2016). Integrative approaches for data analysis in systems biology: Current advances In *2016 II International Young Scientists Forum on Applied Physics and Engineering (YSF)*, pp. 194–198. IEEE.

Fu R & Gong J (2017). Single cell analysis linking ribosomal (r)DNA and rRNA copy numbers to cell size and growth rate provides insights into molecular protistan ecology. *Journal of Eukaryotic Microbiology* **64**(6): 885–896.

Fu S, Ma Y, Yao H, Xu Z, Chen S, Song J & Au KF (2018). IDP-denovo: De novo transcriptome assembly and isoform annotation by hybrid sequencing.

Bioinformatics **34**(13): 2168–2176.

Gaebler-Schwarz S, Davidson A, Assmy P, Chen J, Henjes J, Nöthig EM, Lunau M & Medlin LK (2010). A new cell stage in the haploid-diploid life cycle of the colony-forming haptophyte *Phaeocystis antarctica* and its ecological implications. *Journal of Phycology* **46**(5): 1006–1016.

Gall M, Boyd P, Hall J, Safi K & Chang H (2001). Phytoplankton processes. Part 1: Community structure during the Southern Ocean iron release experiment (SOIREE). *Deep Sea Research Part II: Topical Studies in Oceanography* **48**(11-12): 2551–2570.

Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B & Overington JP (2012). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40**(D1): D1100–D1107.

German National Cohort (GNC) Consortium (2014). The German National Cohort: Aims, study design and organization. *European Journal of Epidemiology* **29**(5): 371–382.

Getoor L & Machanavajjhala A (2012). Entity resolution: Theory, practice & open challenges. *Proceedings of the VLDB Endowment* **5**(12): 2018–2019.

González A, Fillat MF, Bes Mt, Peleato Ml & Sevilla E (2018). The Challenge of Iron Stress in Cyanobacteria In Tiwari A, editor, *Cyanobacteria*, chapter 6. IntechOpen, Rijeka.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N & Regev A (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**(7): 644–652.

Gregory KE, Bird SS, Gross VS, Marur VR, Lazarev AV, Walker WA & Kristal BS (2013). Method development for fecal lipidomics profiling. *Analytical Chemistry* **85**(2): 1114–1123.

Griffith M, Walker JR, Spies NC, Ainscough BJ & Griffith OL (2015). Informatics for RNA sequencing: A web resource for analysis on the cloud. *PLoS Computational Biology* **11**(8): 1–20.

Groussman RD, Parker MS & Armbrust EV (2015). Diversity and evolutionary history of iron metabolism genes in diatoms. *PLoS ONE* **10**(6): e0129081.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N & Regev A (2013). De novo transcript sequence reconstruction from RNA-Seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**(8): 1494–1512.

Hall JL, Ryan JJ, Bray BE, Brown C, Lanfear D, Newby LK, Relling MV, Risch NJ, Roden DM, Shaw SY, Tcheng JE, Tenenbaum J, Wang TN & Weintraub WS (2016). Merging electronic health record data and genomics for cardiovascular research. *Circulation: Cardiovascular Genetics* **9**(2): 193–202.

Hamm C (2000). Architecture, ecology and biogeochemistry of *Phaeocystis* colonies. *Journal of Sea Research* **43**(3-4): 307–315.

Harel A, Bromberg Y, Falkowski PG & Bhattacharya D (2014). Evolutionary history of redox metal-binding domains across the tree of life. *Proceedings of the National Academy of Sciences* **111**(19): 7042–7047.

Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P & Friedman C (2012). Novel data-mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology & Therapeutics* **91**(6): 1010–1021.

Harris JR, Burton P, Knoppers BM, Lindpaintner K, Bledsoe M, Brookes AJ, Budin-Ljosne I, Chisholm R, Cox D, Deschênes M, Fortier I, Hainaut P, Hewitt R, Kaye J, Litton JE, Metspalu A, Ollier B, Palmer LJ, Palotie A, Pasterk M, Perola M, Riegman PH, Van Ommen GJ, Yuille M & Zatloukal K (2012). Toward a roadmap in global biobanking for health. *European Journal of Human Genetics* **20**(11): 1105–1111.

Hasin Y, Seldin M & Lusis A (2017). Multi-omics approaches to disease. *Genome Biology* **18**(1): 83.

Henry S, Hoon S, Hwang M, Lee D & DeVore M (2005). Engineering trade study: extract, transform, load tools for data migration In *2005 IEEE Design Symposium, Systems and Information Engineering*, pp. 1–8. IEEE.

Hewitt R & Watson P (2013). Defining biobank. *Biopreservation and Biobanking* **11**(5): 309–315.

Hoerl AE & Kennard RW (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**(1): 55–67.

Holm S (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**(2): 65–70.

Hornung R & Wright MN (2019). Block forests: Random forests for blocks of clinical and omics covariate data. *BMC Bioinformatics* **20**(1): 1–17.

Hothorn T, Bretz F & Westfall P (2008). Simultaneous inference in general parametric models. *Biometrical Journal* **50**(3): 346–363.

Hough KP, Wilson LS, Trevor JL, Strenkowski JG, Maina N, Kim YI, Spell ML, Wang Y, Chanda D, Dager JR, Sharma NS, Curtiss M, Antony VB, Dransfield MT, Chaplin DD, Steele C, Barnes S, Duncan SR, Prasain JK, Thannickal VJ & Deshane JS (2018). Unique lipid signatures of extracellular vesicles from the airways of asthmatics. *Scientific Reports* **8**(1): 1–16.

Huang S, Chaudhary K & Garmire LX (2017). More is better: Recent progress in multi-omics data integration methods. *Frontiers in Genetics* **8**(84): 1–12.

Human Microbiome Project Consortium (2012). A framework for human microbiome research. *Nature* **486**(7402): 215–221.

Hunter MC, Pozhitkov AE & Noble PA (2017). Accurate predictions of postmortem interval using linear regression analyses of gene meter expression data. *Forensic Science International* **275**: 90–101.

Hutchins DA & Boyd PW (2016). Marine phytoplankton and the changing ocean iron cycle. *Nature Climate Change* **6**(12): 1072–1079.

Iacomino G, Russo P, Stillitano I, Lauria F, Marena P, Ahrens W, De Luca P & Siani A (2016). Circulating microRNAs are deregulated in overweight/obese children: Preliminary results of the I. Family study. *Genes and Nutrition* **11**(1): 3–11.

Islam SM, Kwak D, Kabir MH, Hossain M & Kwak KS (2015). The internet of things for health care: A comprehensive survey. *IEEE Access* **3**: 678–708.

Issak MRR (2014). Transcriptomics of Iron Limitation in *Phaeocystis antarctica* Master's thesis, The American University in Cairo URL: http://dar.aucegypt.edu/handle/10526/3980.

Jacob L, Obozinski G & Vert JP (2009). Group lasso with overlap and graph lasso In *Proceedings of the 26th annual international conference on machine learning*, pp. 433–440.

Ji M, He Q, Han J & Spangler S (2015). Mining strong relevance between heterogeneous entities from unstructured biomedical data. *Data Mining and Knowledge Discovery* **29**(4): 976–998.

Jové M, Naudí A, Portero-Otin M, Cabré R, Rovira-Llopis S, Bañuls C, Rocha M, Hernández-Mijares A, Victor VM & Pamplona R (2014). Plasma lipidomics discloses metabolic syndrome with a specific HDL phenotype. *FASEB Journal* **28**(12): 5163–5171.

Kanehisa M, Furumichi M, Tanabe M, Sato Y & Morishima K (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* **45**(D1): D353–D361.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, Beszteri B, Bidle KD, Cameron CT, Campbell L, Caron DA, Cattolico RA, Collier JL, Coyne K, Davy SK, Deschamps P, Dyhrman ST, Edvardsen B, Gates RD, Gobler CJ, Greenwood SJ, Guida SM, Jacobi JL, Jakobsen KS, James ER, Jenkins B, John U, Johnson MD, Juhl AR, Kamp A, Katz LA, Kiene R, Kudryavtsev A, Leander BS, Lin S, Lovejoy C, Lynn D, Marchetti A, McManus G, Nedelcu AM, Menden-Deuer S, Miceli C, Mock T, Montresor M, Moran MA, Murray S, Nadathur G, Nagai S, Ngam PB, Palenik B, Pawlowski J, Petroni G, Piganeau G, Posewitz MC, Rengefors K, Romano G, Rumpho ME, Rynearson T, Schilling KB, Schroeder DC, Simpson AGB, Slamovits CH, Smith DR, Smith GJ, Smith SR, Sosik HM, Stief P, Theriot E, Twary SN, Umale PE, Vaulot D, Wawrik B, Wheeler GL, Wilson WH, Xu Y, Zingone A & Worden AZ (2014). The marine microbial eukaryote transcriptome sequencing project (MMETSP):

Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biology* **12**(6): e1001889.

Kerfeld CA & Scott KM (2011). Using BLAST to teach "E-value-tionary" concepts. *PLoS Biology* **9**(2): e1001014.

Kindt A, Liebisch G, Clavel T, Haller D, Hörmannsperger G, Yoon H, Kolmeder D, Sigruener A, Krautbauer S, Seeliger C, Ganzha A, Schweizer S, Morisset R, Strowig T, Daniel H, Helm D, Küster B, Krumsiek J & Ecker J (2018). The gut microbiota promotes hepatic fatty acid desaturation and elongation in mice. *Nature Communications* **9**(1): 3760.

Kirpich A, Ainsworth EA, Wedow JM, Newman JR, Michailidis G & McIntyre LM (2018). Variable selection in omics data: A practical evaluation of small sample sizes. *PLoS ONE* **13**(6): 1–19.

Klau S, Jurinovic V, Hornung R, Herold T & Boulesteix AL (2018). Priority-Lasso: A simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC Bioinformatics* **19**(1): 1–14.

Kleppmann M (2017). *Designing Data-Intensive Applications* O'Reilly Media, Inc., Sebastopol, CA.

Kliebenstein D (2012). Exploring the shallow end; estimating information content in transcriptomics studies. *Frontiers in Plant Science* **3**: 213.

Koch F, Beszteri S, Harms L & Trimborn S (2019). The impacts of iron limitation and ocean acidification on the cellular stoichiometry, photophysiology, and transcriptome of *Phaeocystis antarctica*. *Limnology and Oceanography* **64**(1): 357–375.

Koid AE, Liu Z, Terrado R, Jones AC, Caron DA & Heidelberg KB (2014). Comparative transcriptome analysis of four prymnesiophyte algae. *PLoS ONE* **9**(6): e97801.

Lachmann A, Clarke DJ, Torre D, Xie Z & Ma'ayan A (2020). Interoperable RNA-Seq analysis in the cloud. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1863**(6): 194521.

Lachmann A, Xie Z & Ma'Ayan A (2018). Elysium: RNA-seq alignment in the cloud. bioRxiv: 382937.

Lang M, Bischl B & Surmann D (2017). batchtools: Tools for R to work on batch systems. *The Journal of Open Source Software* **2**(10): 135.

Langmead B, Hansen KD & Leek JT (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biology* **11**(8): 1–11.

Li B & Dewey CN (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**(1): 323.

Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R & Dewey CN (2014a). Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology* **15**(12): 1–21.

Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, Juncker AS, Manichanh C, Chen B, Zhang W, Levenez F, Wang J, Xu X, Xiao L, Liang S, Zhang D, Zhang Z, Chen W, Zhao H, Al-Aama JY, Edris S, Yang H, Wang J, Hansen T, Nielsen HB, Brunak S, Kristiansen K, Guarner F, Pedersen O, Doré J, Ehrlich SD, Consortium M, Pons N, Le Chatelier E, Batto JM, Kennedy S, Haimet F, Winogradski Y, Pelletier E, LePaslier D, Artiguenave F, Bruls T, Weissenbach J, Turner K, Parkhill J, Antolin M, Casellas F, Borruel N, Varela E, Torrejon A, Denariaz G, Derrien M, van Hylckama Vlieg JET, Viega P, Oozeer R, Knoll J, Rescigno M, Brechot C, M'Rini C, Mérieux A, Yamada T, Tims S, Zoetendal EG, Kleerebezem M, de Vos WM, Cultrone A, Leclerc M, Juste C, Guedon E, Delorme C, Layec S, Khaci G, van de Guchte M, Vandemeulebrouck G, Jamet A, Dervyn R, Sanchez N, Blottière H, Maguin E, Renault P, Tap J, Mende DR, Bork P & Wang J (2014b). An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology* **32**: 834.

Li YH, Yu CY, Li XX, Zhang P, Tang J, Yang Q, Fu T, Zhang X, Cui X, Tu G, Zhang Y, Li S, Yang F, Sun Q, Qin C, Zeng X, Chen Z, Chen YZ & Zhu F (2018). Therapeutic target database update 2018: Enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Research* **46**(D1): D1121–D1127.

Liu M, McPeek Hinz ER, Matheny ME, Denny JC, Schildcrout JS, Miller RA & Xu H (2012). Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. *Journal of the American Medical Informatics Association* **20**(3): 420–426.

Liu R, AbdulHameed MDM, Kumar K, Yu X, Wallqvist A & Reifman J (2017). Data-driven prediction of adverse drug reactions induced by drug-drug interactions. *BMC Pharmacology and Toxicology* **18**(1): 1–18.

Lommer M, Specht M, Roy AS, Kraemer L, Andreson R, Gutowska MA, Wolf J, Bergner SV, Schilhabel MB, Klostermeier UC, Beiko RG, Rosenstiel P, Hippler M & Laroche J (2012). Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biology* **13**(7): R66.

Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, Lavan P, Weber E, Doak AK, Cote S, Shoichet BK & Urban L (2012). Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486**(7403): 361–367.

Love MI, Anders S, Kim V & Huber W (2015). RNA-Seq workflow: Gene-level exploratory analysis and differential expression. *F1000Research* **4**: 1070.

Love MI, Huber W & Anders S (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology* **15**(12): 550.

Lu Z (2009). Information technology in pharmacovigilance: Benefits, challenges, and future directions from industry perspectives. *Drug, Healthcare and Patient Safety* **1**(1): 35–45.

Magrone T & Jirillo E (2015). Childhood obesity: Immune response and nutritional approaches. *Frontiers in Immunology* **6**: 1–13.

Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM & Stefano GB (2014). Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. *Medical Science Monitor Basic Research* **20**: 138–142.

Marchetti A, Schruth DM, Durkin CA, Parker MS, Kodner RB, Berthiaume CT, Morales R, Allen AE & Armbrust EV (2012). Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proceedings of the National Academy of Sciences* **109**(6): E317–E325.

Marguerat S & Bähler J (2010). RNA-seq: From technology to biology. *Cellular and Molecular Life Sciences* **67**(4): 569–579.

Martin JH, Gordon RM & Fitzwater SE (1990). Iron in Antarctic waters. *Nature* **345**(6271): 156–158.

Martínez O & Humberto Reyes-Valdé M (2008). Defining diversity, specialization, and gene specificity in transcriptomes through information theory. *Proceedings of the National Academy of Sciences* **105**(28): 9709–9714.

Marx V (2013). The big challenges of big data. *Nature* **498**(7453): 255–260.

McMaster C, Liew D, Keith C, Aminian P & Frauman A (2019). A machine-learning algorithm to optimise automated adverse drug reaction detection from clinical coding. *Drug Safety* **42**(6): 721–725.

McTaggart S, Nangle C, Caldwell J, Alvarez-Madrazo S, Colhoun H & Bennie M (2018). Use of text-mining methods to improve efficiency in the calculation of drug exposure to support pharmacoepidemiology studies. *International Journal of Epidemiology* **47**(2): 617–624.

Meier L, Van De Geer S & Bühlmann P (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1): 53–71.

Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM & Culhane AC (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics* **17**(4): 628–641.

Meyer UA (2000). Pharmacogenetics and adverse drug reactions. *Lancet* **356**(9242): 1667–1671.

Misev D & Baumann P (2015). Homogenizing data and metadata retrieval in scientific applications In *Proceedings of the ACM Eighteenth International Workshop on Data Warehousing and OLAP*, DOLAP '15, p. 25–34, New York, NY, USA. Association for Computing Machinery.

Mishra T, Wang M, Metwally AA, Bogu GK, Brooks AW, Bahmani A, Alavi A, Celli A, Higgs E, Dagan-Rosenfeld O, Fay B, Kirkpatrick S, Kellogg R, Gibson M, Wang T, Hunting EM, Mamic P, Ganz AB, Rolnik B, Li X & Snyder MP (2020). Pre-symptomatic detection of COVID-19 from smartwatch data. *Nature Biomedical Engineering* **4**(12): 1208–1220.

Mooney MA, Nigg JT, McWeeney SK & Wilmot B (2014). Functional and genomic context in pathway analysis of GWAS data. *Trends in Genetics* **30**(9): 390–400.

Morandat F, Hill B, Osvald L & Vitek J (2012). Evaluating the design of the R language In Noble J, editor, *ECOOP 2012 – Object-Oriented Programming*, pp. 104–131. Springer, Berlin, Heidelberg.

Morgan XC & Huttenhower C (2012). Chapter 12: Human microbiome analysis. *PLoS Computational Biology* **8**(12): e1002808.

Morrissey J & Bowler C (2012). Iron utilization in marine cyanobacteria and eukaryotic algae. *Frontiers in Microbiology* **3**: 43.

Morrissey J, Sutak R, Paz-Yepes J, Tanaka A, Moustafa A, Veluchamy A, Thomas Y, Botebol H, Bouget FY, McQuaid JB, Tirichine L, Allen AE, Lesuisse E & Bowler C (2015). A novel protein, ubiquitous in marine phytoplankton, concentrates iron at the cell surface and facilitates uptake. *Current Biology* **25**(3): 364–371.

Muhammad S, Ali F & Wrembel R (2017). From conceptual design to performance optimization of ETL workflows: Current state of research and open problems. *The VLDB Journal* **26**(6): 777–801.

Mundra PA, Shaw JE & Meikle PJ (2016). Lipidomic analyses in epidemiology. *International Journal of Epidemiology* **45**(5): 1329–1338.

Nappo A, Iacoviello L, Fraterman A, Gonzalez-Gil EM, Hadjigeorgiou C, Marild S, Molnar D, Moreno LA, Peplies J, Sioen I, Veidebaum T, Siani A & Russo P (2013). High-sensitivity C-reactive protein is a predictive factor of adiposity in children: Results of the identification and prevention of dietary-and lifestyle-induced health effects in children and infants (IDEFICS) study. *Journal of the American Heart Association* **2**(3).

Obozinski G, Jacob L & Vert JP (2011). Group lasso with overlaps: The latent group lasso approach. arXiv: 1110.0413.

Ogata N, Kozaki T, Yokoyama T, Hata T, Iwabuchi K, Martinez O, Reyes-Valdes M, Buettner F, Natarajan K, Casale F, Proserpio V, Scialdone A, Theis F, Saadatpour A, Guo G, Orkin S, Yuan G, Martinez O, Reyes-Valdes M, Herrera-Estrella L, Heil M, Ibarra-Laclette E, Adame-Alvarez R, Martinez O, Ramirez-Chavez E, Molina-Torres J, Ogata N, Yokoyama T, Iwabuchi K, Noori H, Kramer B, Fussenegger M, Luscombe N, Babu M, Yu H, Snyder M, Teichmann S, Gerstein M, Heckel D, Wei P, Zhang J, Dowhan D, Han Y, Moore D, Vee ML, Lecureur V, Moreau A, Stieger B, Fardel O, Loscher W, Klotz U, Zimprich F, Schmidt D, Huang S, Ingber D, Aden D, Fogel A, Plotkin S, Damjanov I, Knowles B, Ohno M, Motojima K, Okano T, Taniguchi A, Mitsuhashi J, Inoue H, Conesa A, Gotz S, Garcia-Gomez J, Terol J, Talon M, Robles M, Marioni J, Mason C, Mane S, Stephens M & Gilad Y (2015). Comparison between the Amount of Environmental Change and the Amount of Transcriptome Change. *PLoS ONE* **10**(12): e0144822.

Orešič M (2009). Bioinformatics and computational approaches applicable to lipidomics. *European Journal of Lipid Science and Technology* **111**(1): 99–106.

Osborn HL & Hook SE (2013). Using transcriptomic profiles in the diatom Phaeodactylum tricornutum to identify and prioritize stressors. *Aquatic Toxicology* **138-139**: 12–25.

Osborne JM, Bernabeu MO, Bruna M, Calderhead B, Cooper J, Dalchau N, Dunn SJ, Fletcher AG, Freeman R, Groen D, Knapp B, McInerny GJ, Mirams GR, Pitt-Francis J, Sengupta B, Wright DW, Yates CA, Gavaghan DJ, Emmott S & Deane C (2014). Ten simple rules for effective computational research. *PLoS Computational Biology* **10**(3): 10–12.

Overington JP, Al-Lazikani B & Hopkins AL (2006). How many drug targets are there? *Nature Reviews Drug Discovery* **5**(12): 993–996.

Pacurariu A, Plueschke K, McGettigan P, Morales DR, Slattery J, Vogl D, Goedecke T, Kurz X & Cave A (2018). Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation. *BMJ Open* **8**(9): e023090.

Paliy O & Shankar V (2016). Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology* **25**(5): 1032–1057.

Panahi B, Frahadian M, Dums JT & Hejazi MA (2019). Integration of cross species RNA-seq meta-analysis and machine-learning models identifies the most important salt stress–responsive pathways in microalga Dunaliella. *Frontiers in Genetics* **10**: 1–12.

Parks MB, Nakov T, Ruck EC, Wickett NJ & Alverson AJ (2018). Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). *American Journal of Botany* **105**(3): 330–347.

Patro R, Duggal G, Love MI, Irizarry RA & Kingsford C (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**(4): 417–419.

Paul S & Chatterjee MK (2020). Data sharing solutions for biobanks for the COVID-19 pandemic. *Biopreservation and Biobanking* **18**(6): 581–586.

Paul S, Gade A & Mallipeddi S (2017). The state of cloud-based biospecimen and biobank data management tools. *Biopreservation and Biobanking* **15**(2): 169–172.

Pearce N (2016). Analysis of matched case-control studies. *BMJ* **352**: i969.

Peiffer-Smadja N, Maatoug R, Lescure FX, D'Ortenzio E, Pineau J & King JR (2020). Machine Learning for COVID-19 needs global collaboration and data-sharing. *Nature Machine Intelligence* **2**(6): 293–294.

Phillips KA, Veenstra DL, Oren E, Lee JK & Sadee W (2001). Potential role of pharmacogenomics in reducing adverse drug reactions: A systematic review. *Journal of the American Medical Association* **286**(18): 2270–2279.

Pigeot I & Ahrens W (2008). Establishment of a pharmacoepidemiological database in Germany: Methodological potential, scientific value and practical limitations. *Pharmacoepidemiology and Drug Safety* **17**(3): 215–223.

Pisa FE, Reinold J, Kollhorst B, Haug U & Schink T (2019). Antidepressants and the risk of traumatic brain injury in the elderly: Differences between individual agents. *Clinical Epidemiology* **11**: 185–196.

Pradhan AD, Manson J, Rifai N, Buring J & Ridker P (2001). C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. *JAMA* **286**(3): 327.

Prjibelski AD, Puglia GD, Antipov D, Bushmanova E, Giordano D, Mikheenko A, Vitale D & Lapidus A (2020). Extending rnaSPAdes functionality for hybrid transcriptome assembly. *BMC Bioinformatics* **21**(Suppl 12): 1–9.

Pullokkaran LJ (2013). Analysis of Data Virtualization & Enterprise Data Standardization in Business Intelligence Master's thesis, Sloan School of Management, Massachusetts Institute of Technology URL: http://hdl.handle.net/1721.1/90703.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD, Wang J, Antolin M, Artiguenave F, Blottiere H, Borruel N, Bruls T, Casellas F, Chervaux C, Cultrone A, Delorme C, Denariaz G, Dervyn R, Forte M, Friss C, Van De Guchte M, Guedon E, Haimet F, Jamet A, Juste C, Kaci G, Kleerebezem M, Knol J, Kristensen M, Layec S, Le Roux K, Leclerc M, Maguin E, Melo Minardi R, Oozeer R, Rescigno M, Sanchez N, Tims S, Torrejon T, Varela E, De Vos W, Winogradsky Y & Zoetendal E (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**(7285): 59–65.

Ramamoorthy C & Wah B (1989). Knowledge and data engineering. *IEEE Transactions on Knowledge and Data Engineering* **1**(1): 9–16.

Rambold G, Yilmaz P, Harjes J, Klaster S, Sanz V, Link A, Glöckner FO & Triebel D (2019). Meta-omics data and collection objects (MOD-CO): A conceptual schema and data model for processing sample data in meta-omics research. *Database* **2019**: 1–13.

Rampelli S, Guenther K, Turroni S, Wolters M, Veidebaum T, Kourides Y, Molnár D, Lissner L, Benitez-Paez A, Sanz Y, Fraterman A, Michels N, Brigidi P, Candela M & Ahrens W (2018). Pre-obese children's dysbiotic gut microbiome and unhealthy diets may predict the development of obesity. *Communications Biology* **1**(1): 222.

Rau A, Marot G & Jaffrézic F (2014). Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics* **15**(1): 1–10.

Reuter JA, Spacek DV & Snyder MP (2015). High-throughput sequencing technologies. *Molecular Cell* **58**(4): 586–597.

Rhee A, Cheong R & Levchenko A (2012). The application of information theory to biochemical signaling systems. *Physical biology* **9**(4): 045011.

Rizkallah MR, Saad R & Aziz RK (2010). The Human Microbiome Project, personalized medicine and the birth of pharmacomicrobiomics. *Current Pharmacogenomics and Personalized Medicine* **8**(3): 182–193.

Robins JM, Gail MH & Lubin JH (1986). More on "Biased selection of controls for case-control analyses of cohort studies". *Biometrics* **42**(2): 293–9.

Rohde H, Qin J, Cui Y, Li D, Loman NJ, Hentschke M, Chen W, Pu F, Peng Y, Li J, Xi F, Li S, Li Y, Zhang Z, Yang X, Zhao M, Wang P, Guan Y, Cen Z, Zhao X, Christner M, Kobbe R, Loos S, Oh J, Yang L, Danchin A, Gao GF, Song Y, Li Y, Yang H, Wang J, Xu J, Pallen MJ, Wang J, Aepfelbacher M & Yang R (2011). Open-source genomic analysis of Shiga-toxin–producing *E. coli* O104:H4. *New England Journal of Medicine* **365**(8): 718–724.

Rubin BE, Wetmore KM, Price MN, Diamond S, Shultzaberger RK, Lowe LC, Curtin G, Arkin AP, Deutschbauer A & Golden SS (2015). The essential gene set of a photosynthetic organism. *Proceedings of the National Academy of Sciences* **112**(48): E6634–E6643.

Rung J & Brazma A (2013). Reuse of public genome-wide gene expression data. *Nature Reviews Genetics* **14**(2): 1–11.

Sanz Y, Romaní-Perez M, Benítez-Páez A, Portune KJ, Brigidi P, Rampelli S, Dinan T, Stanton C, Delzenne N, Blachier F, Neyrinck AM, Beaumont M, Olivares M, Holzer P, Günther K, Wolters M, Ahrens W, Claus SP, Campoy C, Murphy R, Sadler C, Fernández L & van der Kamp JW (2018). Towards microbiome-informed dietary recommendations for promoting metabolic and mental health: Opinion papers of the MyNewGut project. *Clinical Nutrition* **37**(6): 2191–2197.

Saunders G, Ivkovic S, Ghosh R & Yearwood J (2005). Applying anatomical therapeutic chemical (ATC) and critical term ontologies to Australian drug safety data for association rules and adverse event signalling In *Proceedings of the 2005 Australasian Ontology Workshop - Volume 58*, AOW '05, pp. 93–98, Darlinghurst, Australia. Australian Computer Society, Inc.

Savage JH, Lee-Sarwar KA, Sordillo J, Bunyavanich S, Zhou Y, O'Connor G, Sandel M, Bacharier LB, Zeiger R, Sodergren E, Weinstock GM, Gold DR, Weiss ST & Litonjua AA (2018). A prospective microbiome-wide association study of food sensitization and food allergy in early childhood. *Allergy: European Journal of Allergy and Clinical Immunology* **73**(1): 145–152.

Sboner A, Mu XJ, Greenbaum D, Auerbach RK & Gerstein MB (2011). The real cost of sequencing: Higher than you think! *Genome Biology* **12**(8): 125.

Schlenz H, Intemann T, Wolters M, González-Gil EM, Nappo A, Fraterman A, Veidebaum T, Molnar D, Tornaritis M, Sioen I, Mårild S, Iacoviello L & Ahrens W (2014). C-reactive protein reference percentiles among pre-adolescent children in Europe based on the IDEFICS study population. *International Journal of Obesity* **38**(S2): S26–S31.

Schneeweiss S (2010). A basic study design for expedited safety signal evaluation based on electronic healthcare data. *Pharmacoepidemiology and Drug Safety* **19**(8): 858–868.

Schneeweiss S (2018). Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *Clinical Epidemiology* **10**: 771–788.

Schneeweiss S & Avorn J (2005). A review of uses of health care utilization databases for epidemiologic research on therapeutics. *Journal of Clinical Epidemiology* **58**(4): 323–337.

Schoemann V, Becquevort S, Stefels J, Rousseau V & Lancelot C (2005). *Phaeocystis* blooms in the global ocean and their controlling mechanisms: A review. *Journal of Sea Research* **53**(1): 43–66.

Schoffman H, Lis H, Shaked Y & Keren N (2016). Iron–nutrient interactions within phytoplankton. *Frontiers in Plant Science* **7**(1223).

Schöttker B, Saum KU, Muhlack DC, Hoppe LK, Holleczek B & Brenner H (2017). Polypharmacy and mortality: New insights from a large cohort of older adults by detection of effect modification by multi-morbidity and comprehensive correction of confounding by indication. *European Journal of Clinical Pharmacology* **73**(8): 1041–1048.

Shaked Y & Lis H (2012). Disassembling iron availability to phytoplankton. *Frontiers in Microbiology* **3**: 123.

Shannon CE (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**(3): 379–423.

Shi X, Gao W, Chao Sh, Zhang W & Meldrum DR (2013). Monitoring the single-cell stress response of the diatom *Thalassiosira pseudonana* by quantitative real-time reverse transcription-PCR. *Applied and Environmental Microbiology* **79**(6): 1850–1858.

Sielemann K, Hafner A & Pucker B (2020). The reuse of public datasets in the life sciences: Potential risks and rewards. *PeerJ* **8**: e9954.

Smetacek V, De Baar H, Bathmann U, Lochte K & Rutgers Van Der Loeff M (1997). Ecology and biogeochemistry of the Antarctic Circumpolar Current during austral spring: A summary of Southern Ocean JGOFS cruise ANT X/6 of R.V. Polarstern. *Deep Sea Research Part II: Topical Studies in Oceanography* **44**(1-2): 1–21.

Smetacek V, Assmy P & Henjes J (2004). The role of grazing in structuring Southern Ocean pelagic ecosystems and biogeochemical cycles. *Antarctic Science* **16**(4): 541–558.

Smith DR, Arrigo KR, Alderkamp AC & Allen AE (2014a). Massive difference in synonymous substitution rates among mitochondrial, plastid, and nuclear genes of *Phaeocystis* algae. *Molecular Phylogenetics and Evolution* **71**: 36–40.

Smith WO, Ainley DG, Arrigo KR & Dinniman MS (2014b). The oceanography and ecology of the Ross Sea. *Annual Review of Marine Science* **6**(1): 469–487.

Sreedharan VT, Schultheiss SJ, Jean G, Kahles A, Bohnert R, Drewe P, Mudrakarta P, Görnitz N, Zeller G & Rätsch G (2014). Oqtans: The RNA-seq workbench in the cloud for complete and reproducible quantitative transcriptome analysis. *Bioinformatics* **30**(9): 1300–1301.

Srivastava A, Malik L, Sarkar H, Zakeri M, Almodaresi F, Soneson C, Love MI, Kingsford C & Patro R (2020). Alignment and mapping methodology influence transcript abundance estimation. *Genome Biology* **21**(1): 239.

Stekhoven DJ & Buhlmann P (2012). MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**(1): 112–118.

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S & Robinson GE (2015). Big data: Astronomical or genomical? *PLoS Biology* **13**(7): 1–11.

Stewart D, Gibson-Smith K, MacLure K, Mair A, Alonso A, Codina C, Cittadini A, Fernandez-Llimos F, Fleming G, Gennimata D, Gillespie U, Harrison C, Junius-Walker U, Kardas P, Kempen T, Kinnear M, Lewek P, Malva J, McIntosh J, Scullin C & Wiese B (2017). A modified Delphi study to determine the level of consensus across the European Union on the structures, processes and desired outcomes of the management of polypharmacy in older people. *PLoS ONE* **12**(11): e0188348.

Stonebraker M, Abadi D, DeWitt DJ, Madden S, Paulson E, Pavlo A & Rasin A (2010). MapReduce and parallel DBMSs. *Communications of the ACM* **53**(1): 64.

Stonebraker M, Brown P, Poliakov A & Raman S (2011). The architecture of SciDB In *Scientific and Statistical Database Management*, SSDBM'11, pp. 1–16. Springer, Berlin, Heidelberg.

Strickler SR, Bombarely A & Mueller LA (2012). Designing a transcriptome next-generation sequencing project for a nonmodel plant species. *American Journal of Botany* **99**(2): 257–266.

Strzepek RF, Boyd PW & Sunda WG (2019). Photosynthetic adaptation to low iron, light, and temperature in Southern Ocean phytoplankton. *Proceedings of the National Academy of Sciences* **116**(10): 4388–4393.

Strzepek RF & Harrison PJ (2004). Photosynthetic architecture differs in coastal and oceanic diatoms. *Nature* **431**(7009): 689–692.

Strzepek RF, Hunter KA, Frew RD, Harrison PJ & Boyd PW (2012). Iron-light interactions differ in Southern Ocean phytoplankton. *Limnology and Oceanography* **57**(4): 1182–1200.

Strzepek RF, Maldonado MT, a. Hunter K, Frew RD & Boyd PW (2011). Adaptive strategies by Southern Ocean phytoplankton to lessen iron limitation: Uptake of organically complexed iron and reduced cellular iron requirements. *Limnology and Oceanography* **56**(6): 1983–2002.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, a Gillette M, Paulovich A, Pomeroy SL, Golub TR, Lander ES & Mesirov JP (2005). Gene

set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**(43): 15545–15550.

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T & Collins R (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12**(3): e1001779.

Sudmant PH, Alexis MS & Burge CB (2015). Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biology* **16**(1): 1–11.

Suling M & Pigeot I (2012). Signal detection and monitoring based on longitudinal healthcare data. *Pharmaceutics* **4**(4): 607–640.

Suling M, Weber R & Pigeot I (2013). Data mining in pharmacoepidemiological databases In Becker C, Fried R & Kuhnt S, editors, *Robustness and Complex Data Structures*, pp. 351–364. Springer, Berlin, Heidelberg.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, D'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P, Boss E, Bowler C, Follows M, Karp-Boss L, Krzic U, Reynaud EG, Sardet C, Sieracki M & Velayoudon D (2015). Structure and function of the global ocean microbiome. *Science* **348**(6237): 1261359–1261359.

Syme C, Czajkowski S, Shin J, Abrahamowicz M, Leonard G, Perron M, Richer L, Veillette S, Gaudet D, Strug L, Wang Y, Xu H, Taylor G, Paus T, Bennett S & Pausova Z (2016). Glycerophosphocholine metabolites and cardiovascular disease risk factors in adolescents: A cohort study. *Circulation* **134**(21): 1629–1636.

Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P & Kuhn M (2016). STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Research* **44**(D1): D380–D384.

Thamatrakoln K, Korenovska O, Niheu AK & Bidle KD (2012). Whole-genome expression analysis reveals a role for death-related genes in stress acclimation of the diatom *Thalassiosira pseudonana*. *Environmental Microbiology* **14**(1): 67–81.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **73**(3): 273–282.

Trimborn S, Brenneis T, Hoppe C, Laglera L, Norman L, Santos-Echeandía J, Völkner C, Wolf-Gladrow D & Hassler C (2017). Iron sources alter the response of Southern Ocean phytoplankton to ocean acidification. *Marine Ecology Progress Series* **578**: 35–50.

Umemoto K, Goda K, Mitsutake N & Kitsuregawa M (2019). A prescription trend analysis using medical insurance claim big data In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 1928–1939. IEEE.

Ungaro A, Pech N, Martin JF, McCairns RJ, Mévy JP, Chappaz R & Gilles A (2017). Challenges and advances for transcriptome assembly in non-model species. *PLoS ONE* **12**(9): e0185020.

van der Maaten L, Postma E & Herik H (2007). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research* **10**(1).

van der Putten PWH (2010). On Data Mining in Context: Cases, Fusion and Evaluation Ph.D. diss., Leiden Institute of Advanced Computer Science (LIACS), Faculty of Science, Leiden University. URL: https://hdl.handle.net/1887/14600.

Vargas R, Ryder E, Diez-Ewald M, Mosquera J, Durán A, Valero N, Pedreañez A, Peña C & Fernández E (2016). Increased C-reactive protein and decreased interleukin-2 content in serum from obese individuals with or without insulin resistance: Associations with leukocyte count and insulin and adiponectin content. *Diabetes & Metabolic Syndrome* **10**(1): S34–S41.

Velagapudi VR, Hezaveh R, Reigstad CS, Gopalacharyulu P, Yetukuri L, Islam S, Felin J, Perkins R, Borén J, Orešič M & Bäckhed F (2010). The gut microbiota modulates host energy and lipid metabolism in mice. *Journal of Lipid Research* **51**(5): 1101–1112.

Venables WN & Ripley BD (2002). *Modern Applied Statistics with S* Statistics and Computing. Springer, New York, NY, 4th edition.

Verity PG, Brussaard CP, Nejstgaard JC, Leeuwe MA, Lancelot C & Medlin LK (2007). Current understanding of *Phaeocystis* ecology and biogeochemistry, and perspectives for future research. *Biogeochemistry* **83**(1-3): 311–330.

Vigani G, Bashir K, Ishimaru Y, Lehmann M, Casiraghi FM, Nakanishi H, Seki M, Geigenberger P, Zocchi G & Nishizawa NK (2016). Knocking down mitochondrial iron transporter (MIT) reprograms primary and secondary metabolism in rice plants. *Journal of Experimental Botany* **67**(5): 1357–1368.

Vilar S, Harpaz R, Santana L, Uriarte E & Friedman C (2012). Enhancing adverse drug event detection in electronic health records using molecular structure similarity: Application to pancreatitis. *PLoS ONE* **7**(7): e41471.

Wagner A (2005). Energy constraints on the evolution of gene expression. *Molecular Biology and Evolution* **22**(6): 1365–74.

Walsh C, Hu P, Batt J & Santos C (2015). Microarray meta-analysis and cross-platform normalization: Integrative genomics for robust biomarker discovery. *Microarrays* **4**(3): 389–406.

Wang L (2017). Heterogeneous data and big data analytics. *Automatic Control and Information Sciences* **3**(1): 8–15.

Wang S & Gribskov M (2017). Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis. *Bioinformatics* **33**(3): 327–333.

Wang Z, Lachmann A & Ma'ayan A (2019). Mining data and metadata from the gene expression omnibus. *Biophysical Reviews* **11**(1): 103–110.

Wehrens R & Buydens LM (2007). Self- and super-organizing maps in R: The kohonen package. *Journal of Statistical Software* **21**(5): 1–19.

Wehrens R & Kruisselbrink J (2018). Flexible self-Organizing Maps in kohonen 3.0. *Journal of Statistical Software* **87**(7): 1–18.

Wenk MR (2010). Lipidomics: New tools and applications. *Cell* **143**(6): 888–895.

Wickham H (2014). *Advanced R* Routledge, Boca Raton, FL, 1st edition Available from: http://adv-r.had.co.nz/.

Wijmenga C & Zhernakova A (2018). The importance of cohort studies in the post-GWAS era. *Nature Genetics* **50**(3): 322–328.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J & Mons B (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**(1): 160018.

Winnenburg R, Sorbello A & Bodenreider O (2015). Exploring adverse drug events at the class level. *Journal of Biomedical Semantics* **6**(1): 1–10.

Wisecaver JH & Hackett JD (2011). Dinoflagellate genome evolution. *Annual Review of Microbiology* **65**(1): 369–387.

Wold S, Sjöström M & Eriksson L (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58**(2): 109–130.

Wolters M, Ahrens J, Romaní-Pérez M, Watkins C, Sanz Y, Benítez-Páez A, Stanton C & Günther K (2019). Dietary fat, the gut microbiota, and metabolic health – A systematic review conducted within the MyNewGut project. *Clinical Nutrition* **38**(6): 2504–2520.

World Health Organization (2021a). Obesity and Overweight, *WHO* URL: https://www.who.int/en/news-room/fact-sheets/detail/obesity-and-overweight (Accessed on: 2019-07-15).

World Health Organization (2021b). Pharmacovigilance, *WHO* URL: https://www.who.int/teams/regulation-prequalification/regulation-and-safety/pharmacovigilance (Accessed on: 2021-11-15).

Worley B & Powers R (2012). Multivariate analysis in metabolomics. *Current Metabolomics* **1**(1): 92–107.

Yang A, Troup M & Ho JW (2017). Scalability and validation of big data bioinformatics software. *Computational and Structural Biotechnology Journal* **15**: 379–386.

Yang H, Qin C, Li YH, Tao L, Zhou J, Yu CY, Xu F, Chen Z, Zhu F & Chen YZ (2016). Therapeutic target database update 2016: Enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Research* **44**(D1): D1069–D1074.

Yendrek CR, Ainsworth EA & Thimmapuram J (2012). The bench scientist's guide to statistical analysis of RNA-Seq data. *BMC Research Notes* **5**(1): 506.

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, Van Belle C, Chandonia JM, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, Strausberg RL, Frazier M & Venter JC (2007). The Sorcerer II global ocean sampling expedition: Expanding the universe of protein families. *PLoS Biology* **5**(3): e16.

Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, Kraft P & Chatterjee N (2009). Pathway analysis by adaptive combination of p-values. *Genetic Epidemiology* **33**(8): 700–709.

Yuan M & Lin Y (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1): 49–67.

Zambelli F, Mastropasqua F, Picardi E, D'Erchia AM, Pesole G & Pavesi G (2018). RNentropy: an entropy-based tool for the detection of significant variation of gene expression across multiple RNA-Seq experiments. *Nucleic Acids Research* **46**(8): e46.

Zeng Y & Breheny P (2016). Overlapping group logistic regression with applications to genetic pathway selection. *Cancer Informatics* **15**: 179–187.

Zou H (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**(476): 1418–1429.