Following the trail of cellular signatures:

Computational methods for the analysis of molecular high-throughput profiles

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften (Dr. rer. nat.)
der Fakultät für Mathematik und Informatik
der Universität des Saarlandes

von

Tim Kehl

Saarbrücken, 2022

Tag des Kolloquiums:

13. Januar 2023

Dekan der Fakultät:

Prof. Dr. Jürgen Steimle

Prüfungsausschuss:

Vorsitzende des Prüfungsausschusses:

Prof. Dr. Verena Wolf

Berichterstatter:

Prof. Dr. Hans-Peter Lenhof
Prof. Dr. Volkhard Helms

Wissenschaftlicher Mitarbeiter:

Dr. Fabian Kern

ACKNOWLEDGEMENTS

Over the last few years, while writing this thesis, I have received a great deal of support from many people without whom this dissertation would not have been possible. Here, I would like to acknowledge some of them personally.

First of all, I would like to thank my supervisor Prof. Dr. Hans-Peter Lenhof for his consistent support and guidance during the course of my dissertation. Your insights were invaluable for my research and our meetings and discussions benefited this thesis greatly.

Additionally, many thanks should also go to my former colleagues, collaborators, and students I had the pleasure to work with. Thank you for always providing a very pleasant work environment.

Moreover, I am also very grateful to my proof readers Prof. Dr. Andreas Keller, Dr. Patrick Trampert, Dr. Lara Schneider, Lea Eckhart, Nico Gerstner, and Kerstin Lenhof who fought an fierce battle against typos, grammar errors, and unclear formulations.

Furthermore, I would also like to mention my golfmates from the "Golfclub Homburg/Saar Websweiler Hof e.V." who every week requested a short update on the progress of my thesis.

Last but not least, I have to thank my family for their continues support and encouragement throughout the years.

PUBLICATIONS

This dissertation is based on a collection of peer-reviewed journal publications. The main publications discussed in this thesis are listed in the following section. A complete list can be found in Appendix A.

MAIN PUBLICATIONS

Christina Backes, **Tim Kehl**, Daniel Stöckel, Tobias Fehlmann, Lara Schneider, Eckart Meese, Hans-Peter Lenhof, and Andreas Keller. "miRPathDB: a new dictionary on microRNAs and target pathways." In: *Nucleic acids research* (2016), gkw926.

Caroline Diener, Martin Hart, **Tim Kehl**, Stefanie Rheinheimer, Nicole Ludwig, Lena Krammes, Sarah Pawusch, Kerstin Lenhof, Tanja Tänzer, David Schub, et al. "Quantitative and time-resolved miRNA pattern of early human T cell activation." In: *Nucleic Acids Research* (2020).

Nico Gerstner, **Tim Kehl**, Kerstin Lenhof, Lea Eckhart, Lara Schneider, Daniel Stöckel, Christina Backes, Eckart Meese, Andreas Keller, and Hans-Peter Lenhof. "GeneTrail: a framework for the analysis of high-throughput profiles." In: *Frontiers in Molecular Biosciences* (2021), p. 890.

Nico Gerstner, **Tim Kehl**, Kerstin Lenhof, Anne Müller, Carolin Mayer, Lea Eckhart, Nadja Liddy Grammes, Caroline Diener, Martin Hart, Oliver Hahn, et al. "GeneTrail 3: advanced high-throughput enrichment analysis." In: *Nucleic Acids Research* (2020).

Daniel Stöckel, Oliver Müller, **Tim Kehl**, Andreas Gerasch, Christina Backes, Alexander Rurainski, Andreas Keller, Michael Kaufmann, and Hans-Peter Lenhof. "NetworkTrail—a web service for identifying and visualizing deregulated subnetworks." In: *Bioinformatics* 29.13 (2013), pp. 1702–1703.

Daniel Stöckel, **Tim Kehl**, Patrick Trampert, Lara Schneider, Christina Backes, Nicole Ludwig, Andreas Gerasch, Michael Kaufmann, Manfred Gessler, Norbert Graf, et al. "Multi-omics enrichment analysis using the GeneTrail2 web service." In: *Bioinformatics* 32.10 (2016), pp. 1502–1508.

Tim Kehl, Christina Backes, Fabian Kern, Tobias Fehlmann, Nicole Ludwig, Eckart Meese, Hans-Peter Lenhof, and Andreas Keller. "About miRNAs, miRNA seeds, target genes and target pathways." In: *Oncotarget* 8.63 (2017), p. 107167.

Tim Kehl, Fabian Kern, Christina Backes, Tobias Fehlmann, Daniel Stöckel, Eckart Meese, Hans-Peter Lenhof, and Andreas Keller. "miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database." In: *Nucleic acids research* 48.D1 (2020), pp. D142–D147.

Tim Kehl, Lara Schneider, Kathrin Kattler, Daniel Stöckel, Jenny Wegert, Nico Gerstner, Nicole Ludwig, Ute Distler, Markus Schick, Ulrich Keller, et al. "REGGAE: a novel approach for the identification of key transcriptional regulators." In: *Bioinformatics* 1 (2018), p. 8.

Tim Kehl, Lara Schneider, Kathrin Kattler, Daniel Stöckel, Jenny Wegert, Nico Gerstner, Nicole Ludwig, Ute Distler, Stefan Tenzer, Manfred Gessler, et al. "The role of TCF3 as potential master regulator in blastemal Wilms tumors." In: *International journal of cancer* 144.6 (2019), pp. 1432–1443.

Tim Kehl, Lara Schneider, Florian Schmidt, Daniel Stöckel, Nico Gerstner, Christina Backes, Eckart Meese, Andreas Keller, Marcel H Schulz, and Hans-Peter Lenhof. "RegulatorTrail: a web service for the identification of key transcriptional regulators." In: *Nucleic acids research* 45.W1 (2017), W146–W153.

SHORT SUMMARY

Nowadays, molecular profiling techniques, like high-throughput sequencing, microarrays, or mass spectrometry, are routinely applied to generate detailed multi-omics profiles of cells. Since the resulting data sets are high-dimensional and often noisy, powerful and robust computational methods are necessary to enable their analysis.
In this thesis, we present a comprehensive framework of algorithms, tools, and databases to analyze molecular high-throughput profiles. We developed this framework to explore deregulated biological processes involved in the pathogenesis of complex diseases, like cancer, and to detect driving molecules within those processes.
Moreover, we evaluate the capabilities of our tool suite through several case studies that highlight the versatility and potential of our framework. For this purpose, we in particular conducted a detailed study of Wilms' tumors. Here, we identified various regulatory mechanisms, including new potential biomarkers, that might contribute to increased malignancy of the blastemal subtype. These findings could even lead to new therapeutic strategies for Wilms' tumors.
The presented results showcase that our framework is well equipped for the analysis of molecular high-throughput profiles and can help elucidate complex pathogenic mechanisms in cancer and other diseases.

ZUSAMMENFASSUNG

Heutzutage werden molekulare Hochdurchsatzmessverfahren, wie Hochdurchsatzsequenzierung, Microarrays, oder Massenspektrometrie, regelmäßig angewendet, um Zellen im großen Stil und auf verschiedenen molekularen Ebenen zu charakterisieren. Die dabei generierten Datensätze sind in der Regel hochdimensional und oft verrauscht. Daher werden leistungsfähige computergestützte Anwendungen benötigt, um deren Analyse zu ermöglichen.
In dieser Arbeit präsentieren wir eine Reihe von effektiven Algorithmen, Programmen, und Datenbaken für die Analyse von molekularen Hochdurchsetzdatensätzen. Diese Ansätze wurden entwickelt, um deregulierte biologische Prozesse zu untersuchen und in diesen wichtige Schlüsselmoleküle zu identifizieren.
Zusätzlich wurden eine Reihe von Analysen durchgeführt um die verschiedenen Methoden zu evaluieren. Zu diesem Zweck haben wir insbesondere eine Wilmstumor Studie durchgeführt, in der wir verschiedene regulatorische Mechanismen und dazugehörige Biomarker identifizieren konnten, die für die erhöhte Malignität von Wilmstumoren mit blastemreichen Subtyp verantwortlich sein könnten. Diese Erkenntnisse könnten in der Zukunft zu einer verbesserten Behandlung dieser Tumore führen.
Diese Ergebnisse zeigen eindrucksvoll, dass unsere Ansätze in der Lage sind, verschiedene molekulare Hochdurchsatzmessungen auszuwerten und dabei helfen können pathogene Mechanismen im Zusammenhang mit Krebs oder anderen komplexen Krankheiten aufzuklären.

ABSTRACT

Over the last three decades, high-throughput techniques, such as next-generation sequencing, microarrays, or mass spectrometry, have revolutionized biomedical research by enabling scientists to generate detailed molecular profiles of biological samples on a large scale. These profiles are usually complex, high-dimensional, and often prone to technical noise, which makes a manual inspection practically impossible. Hence, powerful computational methods are required that enable the analysis and exploration of these data sets and thereby help researchers to gain novel insights into the underlying biology.
In this thesis, we present a comprehensive collection of algorithms, tools, and databases for the integrative analysis of molecular high-throughput profiles. We developed these tools with two primary goals in mind. The detection of deregulated biological processes in complex diseases, like cancer, and the identification of driving factors within those processes.
Our first contribution in this context are several major extensions of the GeneTrail web service that make it one of the most comprehensive toolboxes for the analysis of deregulated biological processes and signaling pathways. GeneTrail offers a collection of powerful enrichment and network analysis algorithms that can be used to examine genomic, epigenomic, transcriptomic, miRNomic, and proteomic data sets. In addition to approaches for the analysis of individual -omics types, our framework also provides functionality for the integrative analysis of multi-omics data sets, the investigation of time-resolved expression profiles, and the exploration of single-cell experiments.
Besides the analysis of deregulated biological processes, we also focus on the identification of driving factors within those processes, in particular, miRNAs and transcriptional regulators. For miRNAs, we created the miRNA pathway dictionary database miRPathDB, which compiles links between miRNAs, target genes, and target pathways. Furthermore, it provides a variety of tools that help to study associations between them.
For the analysis of transcriptional regulators, we developed REGGAE, a novel algorithm for the identification of key regulators that have a significant impact on deregulated genes, e.g., genes that show large expression differences in a comparison between disease and control samples. To analyze the influence of transcriptional regulators on deregulated biological processes,, we also created the RegulatorTrail web service. In addition to REGGAE, this tool suite compiles a range of powerful algorithms that can be used to identify key regulators in transcriptomic, proteomic, and epigenomic data sets.

Moreover, we evaluate the capabilities of our tool suite through several case studies that highlight the versatility and potential of our framework. In particular, we used our tools to conducted a detailed analysis of a Wilms' tumor data set. Here, we could identify a circuitry of regulatory mechanisms, including new potential biomarkers, that might contribute to the blastemal subtype's increased malignancy, which could potentially lead to new therapeutic strategies for Wilms' tumors.
In summary, we present and evaluate a comprehensive framework of powerful algorithms, tools, and databases to analyze molecular high-throughput profiles. The provided methods are of broad interest to the scientific community and can help to elucidate complex pathogenic mechanisms.

CONTENTS

1 INTRODUCTION 1
1.1 Motivation . . . 2
1.2 Contributions . . . 3
1.3 Thesis outline . . . 5
2 BIOLOGICAL BACKGROUND 9
2.1 Molecular biology of cells . . . 9
2.1.1 The human genome . . . 10
2.1.2 Chromatin structure . . . 11
2.1.3 Genes, gene products, and genomic regions . . . 13
2.1.4 Gene expression . . . 14
2.1.5 Mechanisms of gene regulation . . . 20
2.2 Immune system . . . 25
2.2.1 Mechanisms of the innate immune system . . . 25
2.2.2 Mechanisms of the adaptive immune response . . . 27
2.3 Cancer . . . 30
2.3.1 Cancer development and hallmark properties . . . 30
3 MATERIALS AND METHODS 35
3.1 High-throughput assays . . . 36
3.1.1 High-throughput sequencing . . . 36
3.1.2 Microarrays . . . 46
3.2 Third-party resources . . . 47
3.2.1 Identifier . . . 47
3.2.2 Reference sets . . . 47
3.2.3 Biological processes, signaling pathways, and functional categories . . . 48
3.2.4 Binding sites of transcriptional regulators . . . 48
3.2.5 miRNA-target interactions (MTIs) . . . 50
3.2.6 Drug- and disease-related resources . . . 50
3.3 Fundamentals of hypothesis testing . . . 51
3.3.1 Hypothesis tests . . . 51
3.3.2 Multiple testing correction . . . 53
3.4 Statistical feature selection and group comparison . . . 55
3.4.1 General notation . . . 55
3.4.2 Fold-changes . . . 56
3.4.3 Parametric tests . . . 57
3.4.4 Non-parametric tests . . . 58
3.5 Correlation and distance measures . . . 60
3.5.1 Distance measures . . . 60
3.5.2 Correlation coefficients . . . 61
3.6 Enrichment analysis . . . 63
3.6.1 General structure of an enrichment analysis workflow . . . 63

3.6.2 Over-representation analysis (ORA) 66
3.6.3 Functional class scoring (FCS) 67
3.7 Network analysis 71
3.7.1 Subgraph ILP 71
3.8 Regulator impact analysis 73
3.8.1 RTI-based approaches 73
3.8.2 Motif-based approaches 77
4 GRAVITON 81
4.1 The fundamentals of modern web applications 82
4.1.1 Resources and Uniform Resource Identifier (URI) 82
4.1.2 The Internet Protocol Suite (TCP/IP) 83
4.1.3 Representational State Transfer (REST) and RESTful APIs 84
4.2 Architecture 87
4.2.1 Front end 88
4.2.2 Back end 89
4.2.3 Database layer 92
4.3 General functionality 93
4.3.1 File type validation and sanity checks 93
4.3.2 Identifier mapping 93
5 GENETRAIL WEB SERVICE 97
5.1 Standard enrichment analysis 99
5.1.1 Input data 100
5.1.2 Workflow 101
5.1.3 Visualization of results 105
5.1.4 Integrative analysis 105
5.2 Network analysis 107
5.2.1 Workflow 108
5.3 Epigenomics workflow 109
5.3.1 Chromatin state assignment 109
5.3.2 Identification of chromatin state transitions . . . 110
5.3.3 Enrichment analysis 111
5.4 Single-cell analysis 112
5.4.1 Preprocessing 113
5.4.2 Single-cell enrichment analysis 116
5.4.3 Group comparison and downstream analyses . 119
5.5 Time-series workflow 121
5.5.1 Step 1 - Feature selection 121
5.5.2 Step 2 - Clustering of time-resolved expression measurements 122
5.6 Results 129
5.6.1 Single cell analysis of monocytes from peripheral blood of COVID − 19 patients 129
5.6.2 Time series analysis of the T cell activation process 132
5.7 Summary, discussion, and conclusion 135
6 MIRPATHDB - THE MIRNA PATHWAY DICTIONARY 137

6.1 Materials and Methods . . . 138
6.1.1 miRNA and miRNA candidates . . . 138
6.1.2 Target genes and miRNA-target interactions (MTIs) 138
6.1.3 Target pathways . . . 139
6.1.4 Comparison between miRNAs . . . 139
6.2 Database content . . . 141
6.2.1 miRNA-centric view . . . 141
6.2.2 Pathway-centric view . . . 143
6.3 Analysis tools and example applications . . . 144
6.3.1 Custom pathway heatmaps . . . 144
6.3.2 Maximum targetome coverage analysis . . . 145
6.4 Discussion and Conclusion . . . 146
7 THE REGGAE ALGORITHM 149
7.1 Algorithm . . . 150
7.1.1 Group comparison and feature selection . . . 150
7.1.2 Association analysis . . . 151
7.1.3 Data integration . . . 151
7.1.4 Enrichment analysis . . . 152
7.1.5 Bootstrapping . . . 153
7.1.6 Aggregating REGGAE results . . . 154
7.2 Results . . . 155
7.2.1 Comparison of ER-positive and ER-negative breast cancer cells . . . 155
7.2.2 Perturbation signatures . . . 159
7.3 Discussion and Conclusion . . . 162
8 THE REGULATORTRAIL WEB SERVICE 165
8.1 The RegulatorTrail web service . . . 167
8.1.1 RTI-based workflows . . . 168
8.1.2 Example applications of RTI-based workflows . . . 170
8.1.3 Motif-based workflows . . . 170
8.1.4 Analyzing key regulators in macrophages . . . 172
8.2 Discussion and Conclusion . . . 173
9 TCF3 AS MASTER REGULATOR IN BLASTEMAL WILMS TUMORS 175
9.1 Analysis of gene expression profiles . . . 177
9.1.1 Influential regulators in blastemal Wilms' tumors 177
9.1.2 Influential regulator complexes . . . 181
9.1.3 Analysis of kidney developmental genes . . . 183
9.2 Analysis of histone marks . . . 184
9.2.1 Comparison of histone marks in human ESCs and WT cells . . . 184
9.2.2 Integrative analysis of epigenomic and transcriptomic data . . . 186
9.3 Discussion and Conclusion . . . 187
10 SUMMARY, DISCUSSION, AND CONCLUSION 189
10.1 Summary . . . 190

10.2 Perspectives . . . 192
10.3 Conclusion . . . 193

I APPENDIX 195

A LIST OF PUBLICATIONS 197
A.1 First and joint first author publications . . . 197
A.2 Co-author publications . . . 199

B ADDITIONAL METHODS 201
B.1 Clustering . . . 201
B.1.1 Hierarchical clustering . . . 201
B.1.2 Community clustering . . . 204
B.2 Dimension reduction . . . 206
B.2.1 Principal component analysis (PCA) . . . 206
B.2.2 t-distributed Stochastic Neighbor Embedding (t-SNE) . . . 207
B.2.3 Uniform Manifold Approximation and Projection (UMAP) . . . 208

C TOOLS FOR CLINICAL DECISION SUPPORT 211
C.1 DrugTargetInspector (DTI) . . . 213
C.1.1 Identification of altered drug targets and their impact on biological processes . . . 214
C.1.2 Example application: Colon adenocarcinoma (TCGA-AA-3542) . . . 217
C.2 ClinOmicsTrail[bc] 219
C.2.1 Input data and initial processing steps . . . 219
C.2.2 Identification of tumor characteristics . . . 221
C.2.3 Assessment of therapy options . . . 224
C.2.4 Example application: Hormone receptor-positive, HER2-negative breast tumor (TCGA-BH-A0DT) 226
C.2.5 Discussion and conclusion . . . 227

D PROOFS 229
D.1 Shifted Euclidean distance . . . 229

E ANALYZED DATA SETS, PROCESSING STEPS, AND PARAMETERS 231
E.1 Hepatocellular carcinoma (HCC) - GSE64041 . . . 231
E.1.1 Preprocessing steps . . . 231
E.1.2 Enrichment analyses . . . 231
E.1.3 Network analysis . . . 233
E.2 Comparison of ER+ and ER- breast cancer cell lines . . . 234
E.2.1 General processing . . . 234
E.2.2 Group comparison and feature selection . . . 234
E.2.3 Regulator effect analyses . . . 235
E.3 CD14 monocytes from peripheral blood of COVID-19 patients . . . 238
E.3.1 Data download and preprocessing steps . . . 238
E.3.2 Enrichment analysis . . . 240

E.3.3 Dimension reduction . . . 240
E.4 Time resolved expression profiles of early T cell activation 241
E.4.1 Microarray experiments (GSE136625) . . . 241
E.4.2 Feature selection and clustering . . . 242
E.4.3 Over-representation analysis (ORA) . . . 243
E.5 Expression profiles of B cell lymphomas in Eμ-Myc transgenic mice . . . 244
E.5.1 General processing . . . 244
E.5.2 Group comparison and feature selection . . . 244
E.5.3 Regulator effect analyses . . . 244
E.5.4 Top 25 regulators identified by REGGAE . . . 246
E.6 Knock-out of pluripotency factors in human ESCs . . . 247
E.6.1 General processing . . . 247
E.6.2 Group comparison and feature selection . . . 247
E.6.3 Regulator effect analyses . . . 247
E.7 Key regulators in macrophages . . . 251
E.7.1 General processing . . . 251
E.7.2 TEPIC analysis . . . 251
E.7.3 INVOKE analysis . . . 251
E.8 Wilms' tumor study (Expression profiles) . . . 253
E.8.1 Ethics statement . . . 253
E.8.2 Microarray experiments (GSE98334) . . . 253
E.8.3 Group comparison . . . 255
E.8.4 REGGAE analysis . . . 255
E.8.5 Enrichment analyses . . . 255
E.9 Wilms' tumor study (Histone marks) . . . 257
E.9.1 Data processing . . . 257
E.10 Colon adenocarcinoma analysis (TCGA-AA-3542) . . . 259
E.10.1 Download . . . 259
E.10.2 Data processing . . . 259
E.11 Breast cancer analysis (TCGA-BH-A0DT) . . . 260
E.11.1 Clinical details . . . 260
E.11.2 Download . . . 260
E.11.3 Data processing . . . 260
F SUPPORTED FILE FORMATS 261
F.1 File formats for molecular measurements . . . 261
F.1.1 Plain text files . . . 261
F.1.2 Sparse matrix formats . . . 262
F.2 File formats for genomic intervals . . . 263
F.2.1 BED format . . . 263
F.2.2 GFF format . . . 264
F.2.3 VCF format . . . 265
F.2.4 SEG format . . . 266
F.2.5 IDAT format . . . 266
F.2.6 GMT format . . . 267
G OVERVIEW OF EXTERNAL RESOURCES 269

G.1 Supported organisms 269
G.2 Supported identifier types 270
G.3 Databases 271
G.4 Regulator target-intersections (RTIs) 272

BIBLIOGRAPHY 273

1 INTRODUCTION

The study of cells, their structure, and function has fascinated humans for centuries [343]. Over time, major scientific breakthroughs in cell biology have often been initiated by technological advancements that allowed scientists to get a more comprehensive and detailed view of cells. This development started with the invention of microscopes, which allowed scientists to identify microorganisms like protozoa or bacteria [343]. The term "cell" itself was first used in the book "Micrographia" by Robert Hooke in 1665, where he used it to describe pores in a piece of cork [219].

With better microscopes, biologists discovered that the tissues of animals [481] and plants [471] are composed of cells. Both observations inspired the original formulation of "cell theory", which *inter alia* proposes that cells are the basic building blocks of all organisms [343, 481]. Over the years, scientists were able to refine and expand our knowledge of cells, such that by the end of the 19^{th} century their general structure and many organelles had been described, e.g., nucleus [76], endoplasmatic reticulum [171], or mitochondria [42].

In the 20^{th} century, the research focus shifted from the characterization of cells on a higher level to molecular characterization of individual components and contained macromolecules, mainly DNA and proteins. These efforts, amongst others, led to the identification of DNA as the carrier of the genetic information [26]. Moreover, technologies like X-ray diffraction crystallography allowed determining the structure of proteins [407] and DNA [566].

At the same time, researchers started to study links and relationships between the different macromolecules in cells. These efforts led to several groundbreaking discoveries, such as the "central dogma of molecular biology", which states that the sequence information is passed from DNA to protein [105], or the discovery of the "genetic code" that defines how the DNA is translated into a corresponding amino acid sequence [106].

With the knowledge that the DNA is the carrier of the genetic information, researchers also started to study the DNA sequence itself. Different technologies like "DNA sequencing with chain-terminating inhibitors" [465] made it possible to sequence genomes of small viruses, like the ϕX174 bacteriophage [464], the lambda bacteriophage [463], or the DNA of human mitochondria [21]. These projects showed that it is possible to assemble whole genomes from small fragments and highlighted the potential of cataloging and annotating the complete genomic landscape of organisms.

The success of DNA sequencing and the emergence of new sequencing strategies, such as random shotgun sequencing [360, 423], which allowed to sequence larger genomes, inspired the launch of the Human Genome Project (HGP) in 1990, a huge research program with the goal to sequence and annotate the entire human DNA and thereby to accelerate biomedical research. To this end, the research group collected DNA samples of multiple donors to create a consensus and reference sequence. In 1998, Craig Venter founded the biotechnology company Celera Genomics to initiate a competing project with the goal to generate the first human genome sequence at a faster pace and at a much lower cost than the HGP. Both groups independently published the first versions of the human genome sequence in 2001 [288, 550].
Since then, sequencing technologies have made tremendous progress. The generation of the full human genome sequence by the HGP took multiple years and had an estimated cost up to $1 billion[1] [576]. Nowadays, modern high-throughput sequencing techniques can achieve this task with only a fraction of the time and cost (around $700 [575]).
The success of high-throughput DNA sequencing also inspired the development of further related sequencing protocols that not only allow to determine the DNA sequence of a certain sample (cell mixture), but also epigenetic modifications of the chromatin [500], gene expression [562], or the location of DNA binding proteins on the genome [405].
In the last few years, advancements in single cell isolation techniques such as "laser capture microdissection (LCM)" [140] or "flow-activated cell sorting (FACS)" [65] even allow DNA and RNA sequencing of individual cells instead of cell mixtures [81, 245, 450, 460].

1.1 MOTIVATION

Today, next-generation sequencing (NGS) approaches as well as other high-throughput methods such as mass-spectrometry, or microarrays are routinely applied to profile cells on a large scale, both at bulk and single-cell level.
The wide availability of molecular high-throughput methods also motivated the formation of huge international research projects that try to catalog the molecular markup of healthy as well as diseased cells. Two essential projects that have been the foundation for a large number of significant scientific publications over the last decade are ENCODE [99, 366] and GTEx [314]. The ENCODE project is a resource of various (epi-)genetic data sets for human and mouse. It was

1 The total cost of the HGP was around $2.7 billion. The estimated cost for the generation of the final genome sequence itself was estimated to be in the range of $500 million to $1 billion [576].

created to annotate functional elements in the DNA, to identify and characterize chromatin states in different tissues and cell types, and to locate binding sites of a diverse set of DNA binding proteins. The GTEx project is a repository of tissue-specific gene expression, genetic variants, and epigenetic modifications of 54 tissues and nearly 1,000 individuals. This project aims to study the tissue-specific regulation of gene expression and the influence of (epi-)genetic alterations.
In contrast to ENCODE and GTEx, which have mainly been developed for the pursuit of understanding the general molecular mechanisms, other projects (e.g., [332, 536, 606]) focus more on uncovering pathogenic mechanisms that are responsible for the development of certain diseases. One of them is The Cancer Genome Atlas (TCGA) [536], a comprehensive database containing molecular profiles of more than 20,000 primary cancer samples for 33 different cancer types and matched controls. For each cancer sample, TCGA provides high-throughput measurements of the (epi-)genome, transcriptome, and proteome, as well as clinical information, like treatment strategies or survival times. These efforts already resulted in various discoveries that help to explain the development and progression of different cancer types [382–384].
Due to their popularity and wide availability, high-throughput profiling techniques have even become essential tools in non-research settings, such as forensics [77] or even clinical applications [308].
Nowadays, the biggest challenge no longer seems to be the molecular profiling of cells, but rather the processing of the resulting data to gain new biological insights. Since the size, complexity, and high-dimensionality of the data sets makes a manual inspection impossible, automatic computational resources are required that enable the assessment and exploration of underlying molecular mechanisms.

1.2 CONTRIBUTIONS

This thesis presents algorithms, tools, and databases that facilitate the integrative analysis of molecular high-throughput profiles. The major goal of our tools is to gain novel insights into biological processes that are deregulated in diseases like cancer. Here, we focus on two essential questions: (1) the identification and analysis of deregulated biological processes, and (2) the detection of driving factors in the underlying natural or pathogenic processes. The following paragraphs describe the different scientific contributions discussed in this work.

Our first contributions are several major the GeneTrail web service [176, 177, 510] that make it one of the most powerful toolboxes for the analysis of biological processes available today. GeneTrail offers solutions for two crucial tasks in this context, i.e., enrichment analysis

and network analysis.

Enrichment analysis methods are statistical tests that can be applied to determine if predefined functional categories, e.g., gene sets with a common biological function, are significantly up- or down-regulated in an investigated data set. For this purpose, we built an extensive collection of functional categories from 40 external databases. For the analysis of these categories, GeneTrail offers a comprehensive framework of statistical tests that can be applied to a broad range of application scenarios. This includes (1) classical enrichment analysis of genomic, proteomic, miRnomic, and transcriptomic data sets, (2) integrative analysis of different data types, and (3) specialized workflows for the assessment of epigenetic modifications, time-series data, and single-cell expression profiles.
Besides set-based enrichment analysis approaches, which do not consider interactions between the individual molecular entities, our web service also offers network analysis methods that utilize the graph topology provided by some of the incorporated databases. For this purpose, we combined GeneTrail with our NetworkTrail web service [509] to offer graph-based algorithms that analyze the topology of the provided networks to detect substantially altered paths, subgraphs, or signaling cascades. An advantage of network analysis approaches is that these methods can also highlight key molecules within those networks, e.g., the root of identified subgraphs.

In addition to the detection of deregulated biological processes, a further important task in the analysis of molecular high-throughput profiles is the identification of driving factors within those processes.
Important molecular factors in this context are so-called miRNAs. These small, non-coding RNAs bind to the untranslated region (UTR) on the $3'$ end of their target mRNAs and thereby inhibit the expression of the corresponding gene [392]. Although their modes of action are well studied, some key questions remain. For example, which miRNA regulates a particular pathway or conversely, which biological processes are controlled by a selected miRNA. In order to study both questions, we created miRPathDB [33, 529, 530], a database that offers information about miRNAs, their target genes, and pathways for human and mouse. On top of this, miRPathDB provides several interactive tools for the analysis and comparison of miRNA-pathway interactions.
Further crucial factors in the control of nearly all cellular processes are transcriptional regulators like transcription factors, chromatin modifiers, or co-factors. Alterations in their functions are related to disruptions of critical biological processes [546] and, consequently, have been observed in various diseases [293]. To analyze the influence of transcriptional regulators on deregulated biological processes, we de-

veloped the RegulatorTrail web service [533]. It offers nine methods to investigate the effects of regulators based on transcriptomic, proteomic, or epigenomic data sets. Moreover, we developed REGGAE [531], a novel enrichment-based algorithm for the identification and prioritization of regulators with essential roles in pathological processes.

Throughout this thesis, we demonstrate the capabilities of our tool suite with several analyses that highlight the versatility and potential of our framework. In particular, we used the functionality outlined above to conduct a Wilms' tumor study [532]. Wilms' tumors or nephroblastomas are a type of pediatric kidney tumors. While they generally have a good prognosis, tumors that predominantly consist of blastemal cells after neoadjuvant chemotherapy have a much more unfavorable course of disease than other tumors. Using our tool suite, we were able to detect several properties of blastemal tumors and key transcriptional regulators that might be involved in the increased malignancy of blastemal tumors. These insights could potentially lead to new strategies in the diagnosis and therapy of Wilms' tumors.

An overview of the tools and databases presented in this dissertation is depicted in Figure 1.

1.3 THESIS OUTLINE

This dissertation consists of ten chapters that are structured as described in this section:
In Chapter 2 important concepts of molecular cell biology are introduced. In particular, we present the central biological mechanisms of cells that are relevant for this dissertation. Additionally, we describe how alterations in those mechanisms might be involved in the initiation or progression of cancer.
In Chapter 3, we describe the materials and methods used throughout this thesis. First, the concepts of different high-throughput profiling techniques are explained that were used to gather the data sets analyzed in this thesis. Then we describe the third-party resources that build the data foundation of all our tools. Finally, we introduce important mathematical concepts that are needed in the remaining chapters.
All tools and databases presented in this thesis are implemented as web services using a common framework described in Chapter 4. In particular, we introduce the design and some basic functionality that is shared by all our tools.
In the following chapters, we then describe the individual web services, databases, and algorithms to analyze molecular high-throughput data sets.

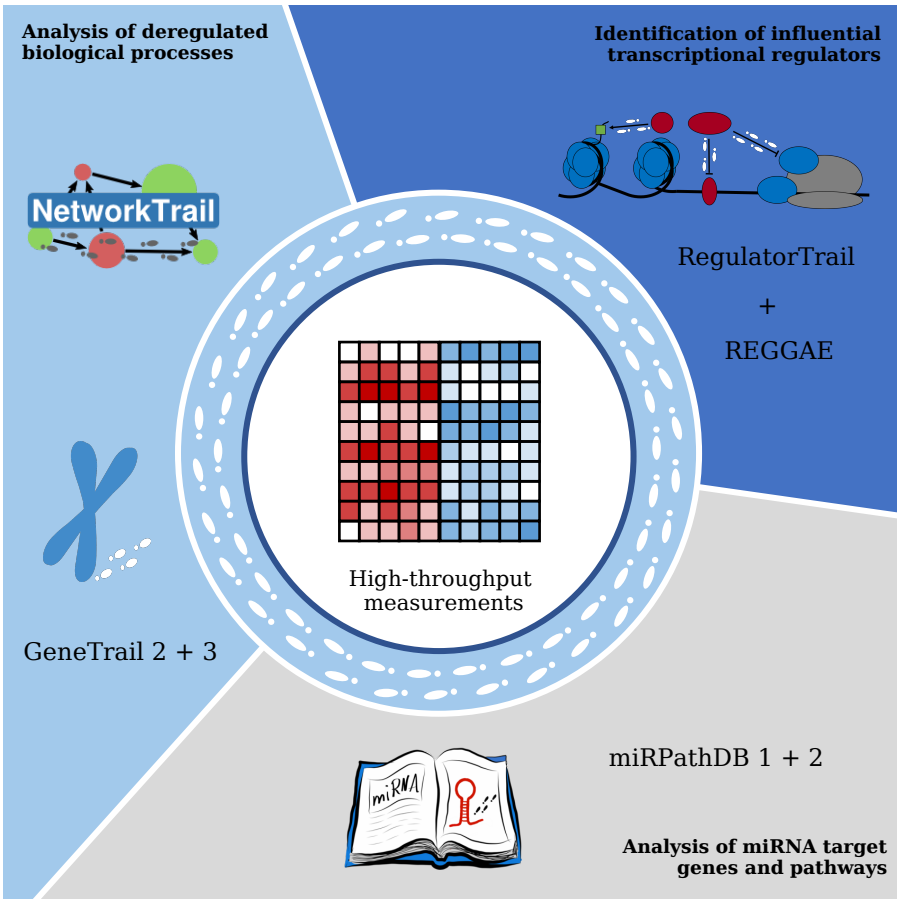


Figure 1: Overview of the algorithms, tools, and databases presented in this thesis.

An overview of the different enrichment and network analysis workflows we created for GeneTrail and NetworkTrail is presented in Chapter 5. In addition to an in-depth discussion of the individual analysis steps, we also show-case the potential of all workflows on real world examples.
Subsequently, we outline how our enrichment analysis framework was used to build miRPathDB and we demonstrate some key aspects of our database (cf. Chapter 6).
In Chapter 7, we introduce, evaluate, and discuss our REGGAE algorithm for the identification of influential regulators.
This is then followed by a description of our RegulatorTrail web service and the different workflows for the investigation of influential regulators in Chapter 8.
In Chapter 9, we describe our study of Wilms tumors, where our tools were used to investigate regulatory mechanisms that may contribute to the increased malignancy of predominantly blastemal tumors.

Finally, in Chapter 10 we summarize and discuss all our contributions and conclude with possible directions of future research.

Author contributions

The research projects described in this work are often based on joint efforts from different research groups across several scientific areas and most of them have already been published in peer-reviewed scientific journals. For this purpose, all chapters contain boxes that reference respective publications and that briefly summarize my contributions.

2 BIOLOGICAL BACKGROUND

In this chapter, we introduce fundamental biological concepts that are needed to understand the remaining chapters of this thesis. First, some basics of molecular cell biology are presented (cf. Section 2.1). Afterward, we discuss several important biological processes. In particular, we provide a brief description of the immune system (cf. Section 2.2) and an overview of processes involved in the development and progression of cancer (cf. Section 2.3).

2.1 MOLECULAR BIOLOGY OF CELLS

Approximation for a 30-year-old person with a height of 1.72m and a weight of 70kg.

The average human body consists of approximately 37.2 trillion cells [59] with various specialized functions.[1] All human cells in an individual's body originate from a single cell, the zygote, which is formed via the fusion of an egg cell and a sperm cell. After fertilization, the zygote proliferates and divides into many daughter cells, which then form the so-called blastocyst [601]. The inner wall of this structure consists of embryonic stem cells (ESCs) that have the ability to transform into any of the three germ layers of the embryonic development (ectoderm, endoderm, mesoderm) and subsequently into any adult cell type [601].
Although all cells of human tissues originate from the same zygote and with some exceptions, e.g., mature erythrocytes[2], also contain the same genetic information, they can fulfill a large spectrum of different functions. This diversity is mainly defined by the number of active genes and the number of functional gene products, such as proteins and non-coding RNAs (cf. Section 2.1.3), but can also be influenced by external stimuli and nutrient supply. In the following sections, we describe the structure of the human genome and various molecular mechanisms that control which genes are active and at which rate gene products are produced.

Most of the information in the following sections is based on the book 'Molecular Biology of the Cell' by Bruce Alberts and colleagues [14]. Other sources are indicated as additional citations.

1 Here, only human cells are considered, not the microbiome.
2 Mature erythrocytes or red blood cells have no nucleus and no DNA.

2.1.1 *The human genome*

The human genome is a set of long double-stranded deoxyribonucleic acid (DNA) molecules that encode the complete hereditary information. It consists of 23 chromosome pairs that are located in the cell nuclei. One copy of each chromosome is inherited from the father and one from the mother. Additionally, a small amount of DNA can also be found in mitochondria. This mitochondrial DNA is solely inherited from the mother.

The double-stranded DNA molecules consist of two unbranched and complementary sequences composed of the same four monomers, called nucleotides. Each nucleotide is composed of a five-carbon sugar in the form of 2-deoxyribose, a phosphate residue, and one out of four nucleobases: adenine (A), cytosine (C), guanine (G), and thymine (T). These nucleotides are covalently bound through the sugar-phosphates and form a single DNA strand. In particular, the fifth carbon of the 2-deoxyribose of one nucleotide is connected to the third carbon of the 2-deoxyribose of the neighboring nucleotide via a phosphate group. This orientation also defines how a strand of DNA is described in the literature, i.e., from the 5′ end to the 3′ end. Finally, two complementary and antiparallel DNA strands are coiled around each other and generate a DNA double helix. The helix structure is stabilized via hydrogen bonding between complementary bases in the two strands. In particular, A always forms base pairs with T and C with G. In Figure 2, the individual components of the DNA double-strand are depicted.

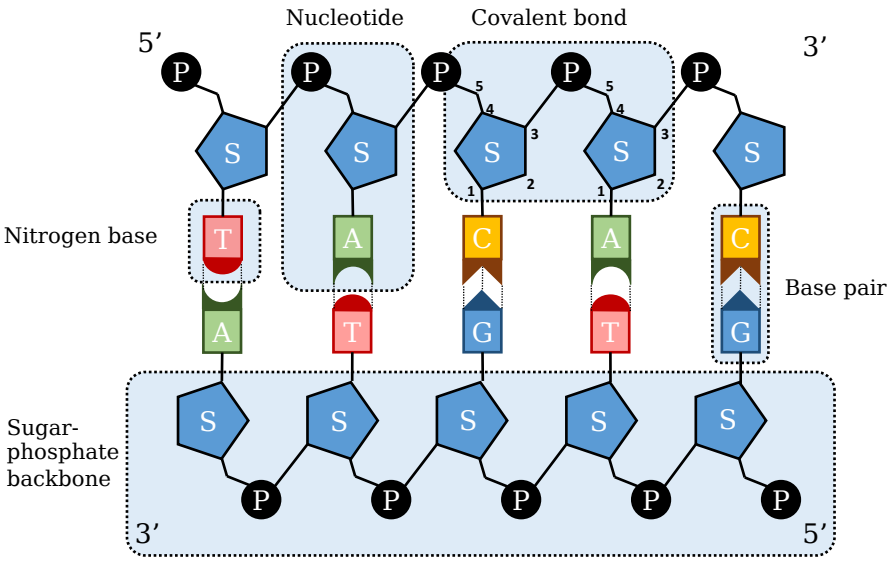


Figure 2: DNA double strand consisting of two complementary, antiparallel DNA chains. Individual components are marked in blue.

2.1.2 *Chromatin structure*

The human DNA contains specific regions, called genes that constitute the blueprint to synthesize thousands of macromolecules, i.e., proteins and non-coding RNA. However, since cells typically only require a subset of the encoded information, only some parts of the DNA are accessible at a given time. The remaining regions are typically condensed into more compact structures. To this end, the DNA double-strand is wrapped around proteins that help organize and condense the DNA. The resulting protein-DNA complex is called chromatin.
The chromatin complex is defined by small subunits, called nucleosomes. Each nucleosome consists of 147bp of DNA that is 1.7 times wrapped around a protein polymer with two copies of each of the following histone proteins: H2A, H2B, H3, and H4 (cf. Figure 3 C). Consecutive nucleosomes are connected by DNA linker regions with a length of 10-80 bp and stabilized by an additional histone (H1) [147]. The arrangement of the nucleosomes on the DNA defines the so called chromatin structure that can range from a very loose representation (euchromatin) to very compact ones (heterochromatin). An overview of the chromatin structure is shown in Figure 3 A+B.

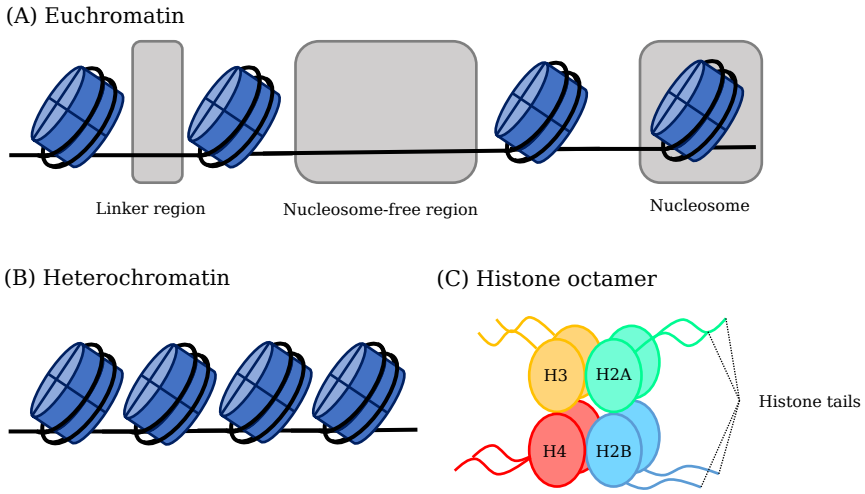


Figure 3: Chromatin structure. (A) Loosely packed euchromatin including a nucleosome-free region. (B) Densely packed heterochromatin. (C) Building blocks of the histone octamer.

2.1.2.1 *Histone modifications*

A shared feature among the four core histones is that they have an unstructured domain at their N-terminal end, which is called the "tail" [60]. This tail is often modified by reversible post-translational modifications (PTMs)[37, 60] (cf. Figure 4). These dynamic modifications change how the histones within and between nucleosomes interact and thereby influence the chromatin structure [60].
Regularly occurring PTMs are methylations, acetylations, phosphorylations, and ubiquitinations. Although several amino acids can be modified, they mainly seem to occur at lysines and arginines [37]. On the one hand, these histone modifications influence the chromatin structure and thereby make the DNA more or less accessible to other proteins. On the other hand, specific combinations also seem to facilitate the interaction of different proteins with the DNA and even large protein complexes that regulate various biological processes, most notably the expression of genes.
We use the following commonly used trilateral notation throughout this thesis to denote specific histone modifications, e.g., H3K4me3. The first part indicates which histone is affected (H3), the second part denotes the one-letter code of the modified amino acid and the position (K4 ⇒ fourth lysine), and the last part specifies the type of modification. (me3 ⇒ trimethylation).

(A) Chromatin modifications

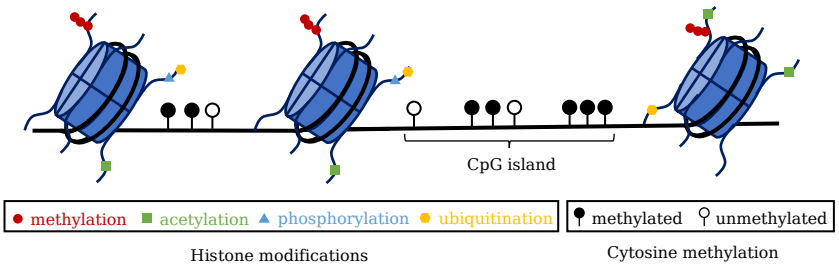


Figure 4: Overview of possible chromatin modifications, i.e., post-translational modification of histones and cytosine methylations.

2.1.2.2 *DNA methylation*

For mammals, non-CpG methylations are mainly found in stem cells and seem to disappear in mature tissues [310, 428].

Another factor that can influence the chromatin structure and the accessibility of the DNA are methylations of nucleobases[3], in particular cytosines [367]. While cytosine methylations can occur in different sequence contexts, it mainly occurs in regions where cytosine (C) is succeeded by guanine (G), i.e., CpG (or CG) sites [310, 428]. Genomic regions with an enrichment of CpG sites are often referred to as CpG islands (CGI).

3 DNA methylation has been identified for adenine and cytosine bases [431, 587].

Similar to histone modifications, DNA methylation, can also influence the accessibility of the DNA and is an essential factor in the control of developmental processes, tissue-specific gene activity, and X chromosome inactivation [367]. While patterns of histone marks can change rapidly, DNA methylation patterns seem to be more stable chromatin modifications [411]. They even have been shown to be heritable [215].

2.1.3 *Genes, gene products, and genomic regions*

The genome of living organisms contains the information to produce other macromolecules, i.e., RNAs and proteins, needed to control the development, characteristics, and functions of a cell. This information is encoded in special genomic regions, called genes. In this section, we first briefly introduce different gene products and genomic regions associated with a gene. The synthesis and regulation of the gene products are then described in subsequent sections.
Each gene represents a template to produce a complementary RNA molecule (cf. Section 2.1.4.1). Like DNA, RNA or ribonucleic acid is composed of covalently bound nucleotides that form a single strand. However, RNA differs from DNA in two key points. Namely, the sugar-phosphate backbone is formed using ribose instead of deoxyribose, and thymine is replaced by uracil. In general there are two distinct classes of RNA molecules: (1) coding RNA molecules that represent a blueprint to create proteins (cf. Section 2.1.4.2) and (2) functional RNAs, such as miRNAs or tRNAs (cf. Section 2.1.4.3 + 2.1.4.2), that are actively involved in cellular processes .
Proteins are polymers consisting of unbranched sequences of monomers, called amino acids, that are linked by covalent peptide bonds. Thereby, each amino acid sequence or polypeptide is encoded by a corresponding RNA (cf. Section 2.1.4.2). After their synthesis, the polypeptide chain folds into a three-dimensional structure that defines the function of a protein (cf. Section 2.1.4.2).
In addition to genes, the genome of most organisms also contains several gene regulatory regions (promoter, enhancer, silencer, and insulator) that are used to control the amount of synthesized gene product (cf. Section 2.1.4). The structure and function of the different regulatory regions are further discussed in Section 2.1.5.3. An overview of the most relevant genomic regions associated with a gene is shown in Figure 5.

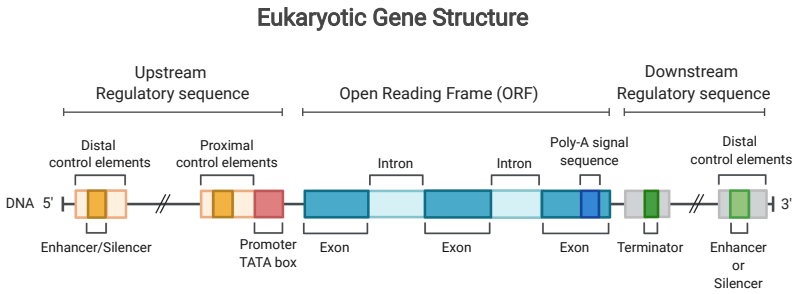


Figure 5: Overview of the genomic regions associated with a gene. This figure was adapted from "Eukaryotic and Prokaryotic Gene Structure", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates.

2.1.4 *Gene expression*

Gene expression is the process of synthesizing functional gene products, such as proteins or non-coding RNAs, based on the sequence of nucleotides that belong to the encoding gene. For the different classes of gene products, distinct processing steps are required. In the following paragraphs, we briefly describe the essential steps needed to synthesize proteins and miRNAs in eukaryotic cells (cf. Figure 6).

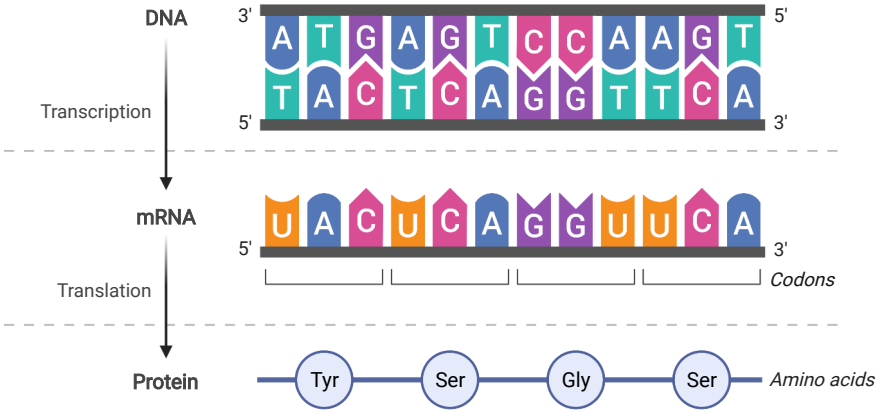


Figure 6: General overview of the flow of genetic information from a gene to a functional gene product. The DNA sequence is transcribed to the corresponding RNA and then translated to the encoded protein. This figure was adapted from "Central Dogma", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates.

2.1.4.1 *Transcription*

The first step in the gene expression process for all functional gene products is transcription. In this process, the DNA sequence of the gene is copied or transcribed into a complementary RNA strand. Transcription is catalyzed by an RNA polymerase and further regulated by transcription factors and cofactors (cf. Section 2.1.5, Figure 7).

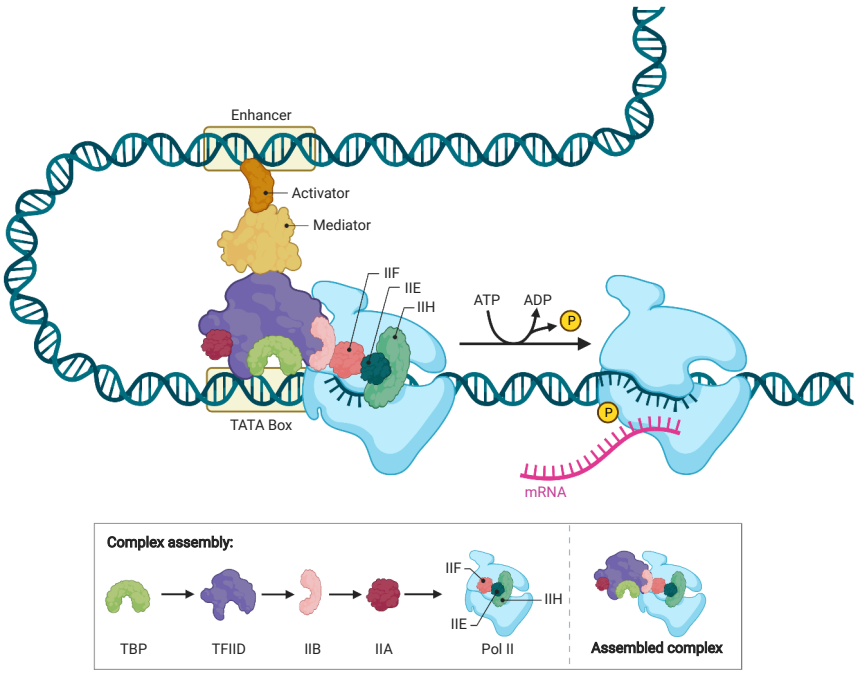


Figure 7: Overview of the transcription initiation process for a specific gene. The (pre-)initiation complex consisting of the polymerase (Pol II) and general transcription factors (TBP, TFIID, IIB, IIA,IIF,IIE,IIIH) is assembled at the promoter region of a gene. The transcription is additionally controlled by an activator protein bound to an associated enhancer region that is looped into the proximity of the polymerase complex. The activation signal from the enhancer is transmitted by the mediator complex (cf. Section 2.1.5.3). This figure was adapted from "Eukaryotic Gene Regulation - Transcriptional Initiation", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates.

The transcription is initiated at the promoter region of a gene, where the RNA polymerase binds with the help of general transcription factors to specific sequences on the DNA. Here, the (pre-)initiation complex consisting of the RNA polymerase and further transcription factors and cofactors is assembled. This complex first separates the two complementary DNA strands and creates a transcription bubble. Additionally, the RNA polymerase is guided to the transcription start site (TSS), where it binds to the single-stranded DNA.

The RNA polymerase now reads the DNA sequence base-by-base from 5′ to 3′ end and at the same time synthesizes a complementary RNA strand (3′ to 5′). After this elongation process has started, the general transcription factors are released to assemble a new (pre-)initiation complex. The RNA polymerase continues to traverse over the template DNA and further extends the complementary RNA strand until the transcription end site (TES) is reached, and the elongation step terminates. Here, the RNA is cleaved, and the polymerase is released from the DNA.

Eukaryotes usually have three different RNA polymerases that catalyze the synthesis of specific RNA molecules. Both mRNAs, which act as templates for protein production, and miRNAs are synthesized by RNA polymerase II. The resulting transcripts for both RNA classes are additionally modified during or after the actual transcription. For mRNA, this entails that (1) a modified nucleotide is added to the 5′ end (capping), (2) the 3′ end is extended with multiple adenines (polyadenylation), and (3) intron sequences (cf. Figure 5) are removed from the sequence (splicing). For miRNAs, several synthesis pathways exist that require different processing steps (cf. Section 2.1.4.3).

2.1.4.2 *Translation and protein folding*

After the transcription process is finished, the resulting mRNA molecule is transported from the nucleus to the cytoplasm of the cell. Here, the nucleotide sequence of the mRNA is translated into a corresponding amino acid sequence of the protein. This process is catalyzed by RNA-protein complexes called ribosomes, that are composed of two subunits (small and large). The ribosomes map three consecutive and non-overlapping nucleotides (codon) of the mRNA sequence to one specific amino acid. To this end, they utilize tRNA molecules that have complementary bases (anticodon) and carry the corresponding amino acid.
The translation is initiated by the small ribosomal subunit that attaches to the start codon of the mRNA (AUG) via complementary base pairs of a tRNA molecule. In the next step, the large ribosomal subunit also attaches and completes the ribosome. The ribosome then traverses the RNA molecule (5′ to 3′), recruits for every codon an associated tRNA, and attaches the corresponding amino acid to the C terminal end of the polypeptide chain. As soon as one of the stop codons (UAA, UAG, UGA) is reached, the translation is stopped, and the amino acid sequence is released. An overview of this process is shown in Figure 8.

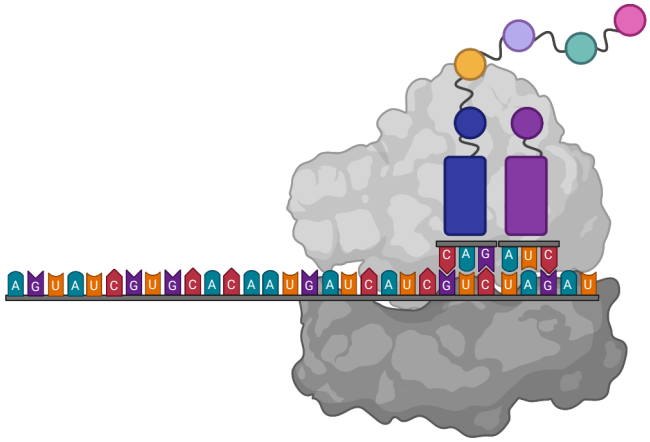


Figure 8: Overview of the translation process. The ribosome traverses a template mRNA and continuously attaches a further amino acid for every three bases to the peptide chain. This figure was created using BioRender.com.

After the translation process, we obtain a chain of amino acids that define the primary structure of the protein. Interactions between the amino acid then cause the protein to fold into a compact three-dimensional conformation. During this process, the protein adopts an ener-

getically favorable structure, i.e., its native conformation. The different levels of the protein structure are depicted in Figure 9.

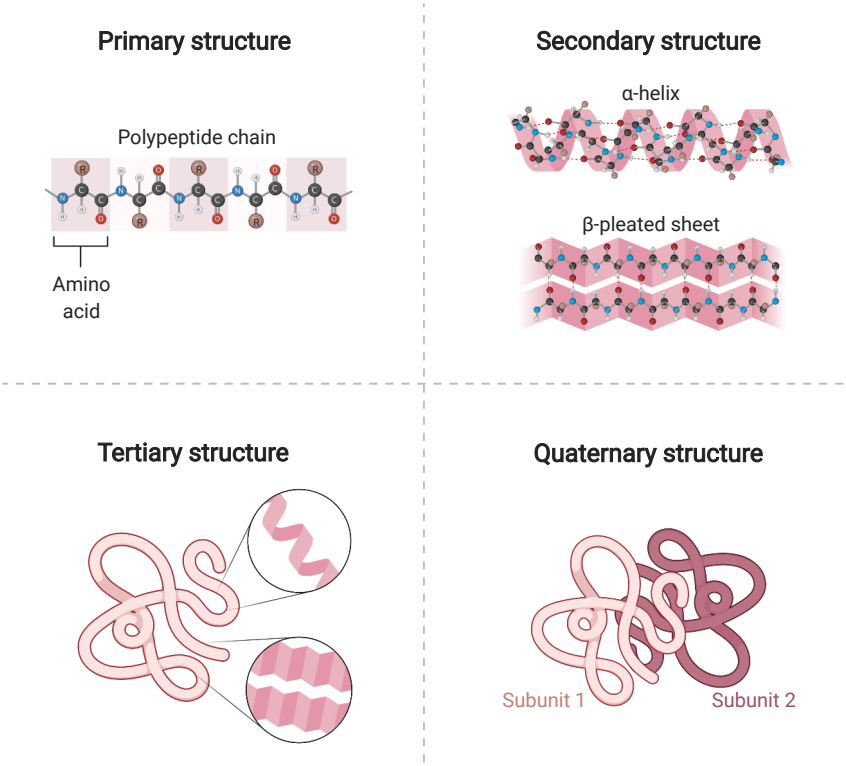


Figure 9: Overview of the different levels of the protein structure. This figure was adapted from "Protein Structure ", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates.

In some cases, the protein folding is also assisted by proteins, called chaperones. The proteins aid in the folding process and help to prevent misfolding and the creation of incorrect conformations.

2.1.4.3 *miRNA biogenesis*

MicroRNAs (miRNAs) are regulatory RNAs with a length between 21-25 nucleotides that fulfill key functions in gene regulation [392, 555]. Each miRNA can repress the expression of a specific set of target genes. To this end, the miRNA binds to the 3′ untranslated region (UTR) of a particular mRNA, which either promotes the degradation of corresponding mRNA molecules or inhibits their translation (cf. Section 2.1.5.4). Here, we briefly describe different ways in which miRNAs can be synthesized [392], i.e., the canonical way and several non-canonical ones.

In the canonical way, genes that encode the miRNA are transcribed from the DNA and then further processed. Here, one typically gene encodes for two mature miRNAs that are named according to their position in the gene body (3p and 5p, cf. Figure 5C). Apart from that, some miRNAs form clusters in the genome, where sequences of multiple miRNAs are grouped and transcribed as one long RNA transcript [522].
After transcription, the primary miRNA transcript (pri-miRNA) is further processed. First, the transcript is processed by a microprocessor protein complex consisting of the two enzymes DGCR8 and Drosha to create the characteristic hairpin structure (pre-miRNA) [117] (cf. Figure 10). The pre-miRNA is then exported to the cytoplasm by a protein called Exportin-5 and subsequently processed by the endonuclease Dicer to create the two mature miRNAs [117, 395].
Additionally, there are several non-canonical ways either independent of Drosha and DGCR8 or independent of Dicer [27, 392]. One example for DGCR8/Drosha independent miRNA biogenesis are mirtrons that are produced from introns of mRNAs via splicing [27, 453]. Examples for Dicer independent biogenesis are miRNAs generated from short hairpin RNAs (shRNAs) by Drosha.

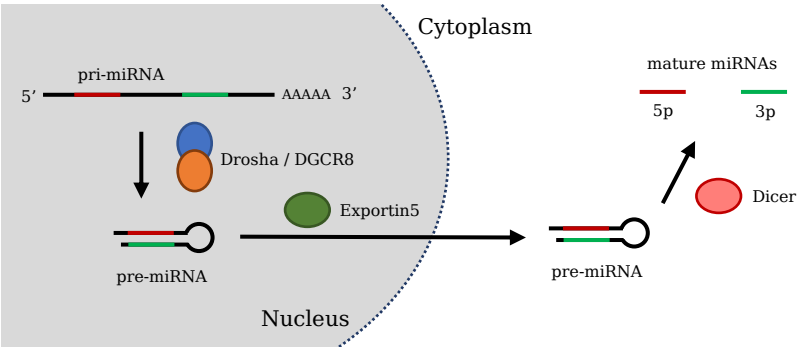


Figure 10: Overview of the canonical miRNA biogenesis pathway. First, the pri-miRNA transcript is processed by Drosha and DGCR8. The resulting stem-loop is then exported to the cytoplasm. Here, Dicer cuts the stem-loop to create two mature miRNAs, which are named according to their position in the primary transcript.

2.1.5 *Mechanisms of gene regulation*

Gene expression can be controlled at nearly all processing steps in the flow of genetic information from DNA to functional gene products (cf. Figure 6). For most genes, the final expression rate is defined by the interplay of different types of regulation. In the following sections, we briefly describe different regulatory mechanisms relevant to the work presented in this thesis. In particular, we focus on transcriptional regulation and gene regulation by miRNAs.

2.1.5.1 *Transcriptional regulation via chromatin remodeling*

Epigenetic modifications of the chromatin constitute the first layer of gene expression control. They define the structure of the chromatin and, hence, regulate if the RNA polymerase complex is able to access the DNA and to initiate the transcription of a gene. Consequently, epigenetic modifications of histones or cytosines have important regulatory roles in many biological processes, e.g., cell development [556, 607].
The chromatin structure in eukaryotic cells is primarily orchestrated by different families of proteins or protein complexes that either modify the chromatin modification patterns [244, 458], or that remodel the chromatin structure by adding, rearranging, or removing nucleosomes [458].

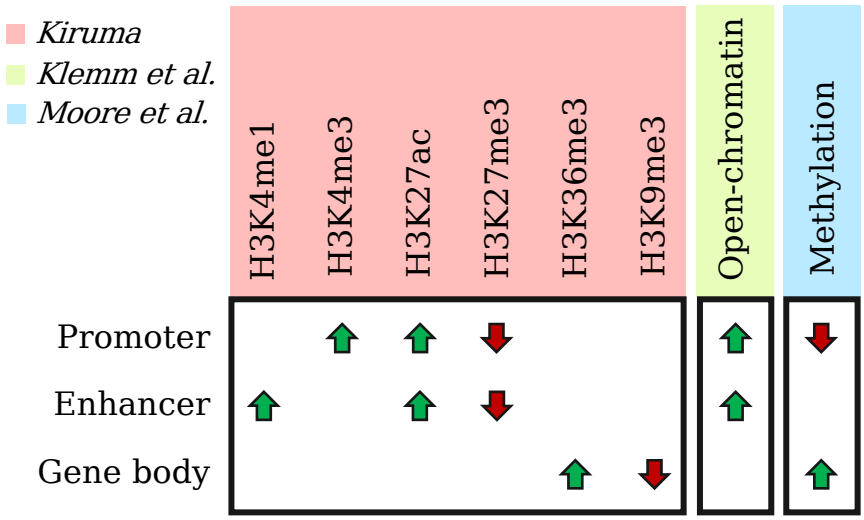


Figure 11: Overview of the epigenetic marks and their association with the activity of a gene. Green arrows indicate a positive, and red arrows a negative association. The information depicted in this plot was extracted from different scientific publications [268, 270, 367].

Regulatory effects of histone modification patterns

Post-translational modifications of histones have two major modes of action [37]. On the one hand, specific modifications influence the packing of nucleosomes and, hence, the accessibility of specific genomic regions. On the other hand, combinations of histone marks define specific chromatin states that facilitate or impede interactions of regulators and DNA [37, 139]. For example, distinct combinations of histone marks in the different regulatory regions of a gene are associated with different stages of activity [139, 268]. An overview of different histone modifications and their association with gene activity are depicted in Figures 11 and 12.

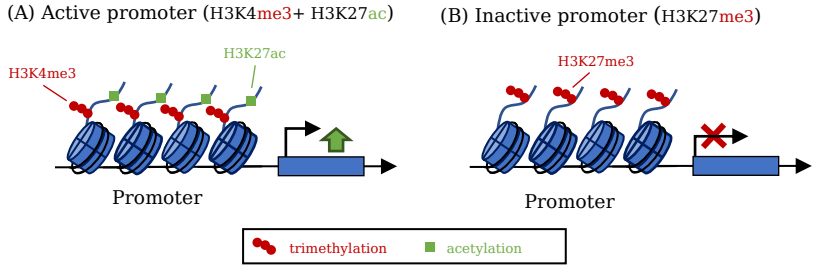


Figure 12: Regulatory effects of histone modification patterns. (A) Combinations of different marks (H3K4me3 and H3K27ac) in the promoter regions indicate an active transcription state. (B) Patterns of H3K27me3 marks in the promoter indicates a silenced state.

Regulatory effects of DNA methylation patterns

Similar to the modification of histones, DNA methylation patterns can influence the expression of genes in two ways.
First of all, the methylation of transcription factor binding sites can prevent them from binding to the DNA [367]. Depending on the affected genomic region and regulator, this can have an activating and a repressing effect on the expression of a gene [247](cf. Figure 11). In general, a high degree of methylation in the promoter region of a gene is associated with transcriptional silencing, while high methylation levels in repressive elements can inhibit their effect [247] (cf. Figure 13).
Additionally, methylation patterns can constitute signals that promote the recruitment of specific protein complexes that then, in turn, mediate the gene expression process [206, 367].

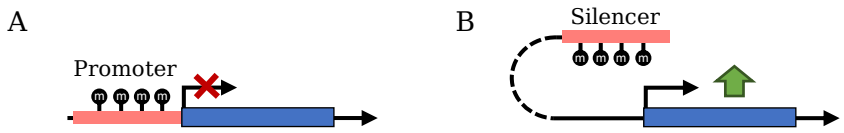


Figure 13: Regulatory effects of DNA methylation patterns. (A) Methylation of promoter regions can repress or silence the expression of genes. (B) Methylation of silencers can reduce/negate their effect.

2.1.5.2 *Gene regulation via transcription factors and cofactors*

The expression of a gene is not only controlled by the accessibility of the DNA to the transcription machinery but also by proteins called transcription factors (TFs). These proteins bind to regulatory regions of a gene and influence the rate at which this gene is transcribed.
In the following paragraphs, we describe the different types of regulatory regions and essential properties of TFs that allow them to influence the expression of their target genes.

2.1.5.3 *Gene regulatory regions (GRR)*

Gene regulatory regions (or cis-regulatory regions) are intervals in the genome that are involved in gene expression or the regulation of this process (cf. Figure 5A). They are enriched with binding sites of transcription regulators that can control the expression of associated genes [438]. In the following paragraphs, we briefly describe some of the core regulatory regions.

CORE PROMOTER The core promoter is defined as the regions around the transcriptional start sites (TSS) of a gene [438, 494]. It contains various binding motifs of general transcription factors that facilitate the formation of the pre-initiation complex, e.g., the TATA box that can be found in around 24% of human promoters [494, 591].

PROXIMAL PROMOTER The proximal promoter is located upstream of the core promoter. It also contains binding motifs of transcriptional regulators. Besides binding sites of activators and repressors, they also often contain tethering elements that enable long-range interactions between the promoter and other regulatory elements, such as distal enhancers [83, 438].

ENHANCER Enhancers are regions enriched with transcription factor motifs that increase the rate at which a specific gene is transcribed [340, 438]. Unlike promoters that are always located near the respective TSS, enhancers can also be located hundreds of kilobases away from the regulated gene [340]. To unfold their effect, these distal enhancers are looped into the close proximity of the corresponding pro-

moter region (cf. Figure 7). This loop formation is mediated by a large protein complex, called the Mediator complex, which transmits the signal of bound regulators to the transcription machinery [503].

SILENCER Silencers are the counterpart to enhancer regions as they generally have a negative effect on the expression of their associated target gene [438].

INSULATOR Insulators are genomic regions that can block interactions between a gene and distal regulatory elements [340]. In this context, they have two main functions. First of all, they can block the communication between a promoter and distal regulatory regions, such as enhancer or silencer, by creating a physical barrier that interferes with the loop formation [173]. Additionally, they can prevent euchromatin domains from being silenced by blocking the spread of repressive heterochromatin [340].

Transcription factors are characterized by their DNA-binding domains

In general, each transcription factor has a specific DNA-binding domain that allows them to recognize certain DNA motifs. The occurrence of these motifs in the regulatory regions of a gene is one important factor that defines if the expression of a gene is influenced by a transcription factor.

Transcription factors act in large protein complexes

Most eukaryotic TFs bind cooperatively to the DNA and assemble complexes with other transcription factors or cofactors. The composition of such complexes defines if and how each regulator contributes to the final function and subsequently how a target gene is regulated.

The activity of each gene is defined by the combined effect of all regulators

If a particular gene is expressed and at which rate is defined by the combined effect of all involved regulators and associated complexes on the RNA polymerase (cf. Figure 14). In this context, the different regulators can fulfill various functions that regulate every step of the transcription process. For example, they can enable, block, or influence the (pre-)initiation complex assembly or the rate at which the polymerase is released from the start site. Additionally, regulators can also affect the expression of a gene by recruiting other protein complexes, such as repressor complexes.

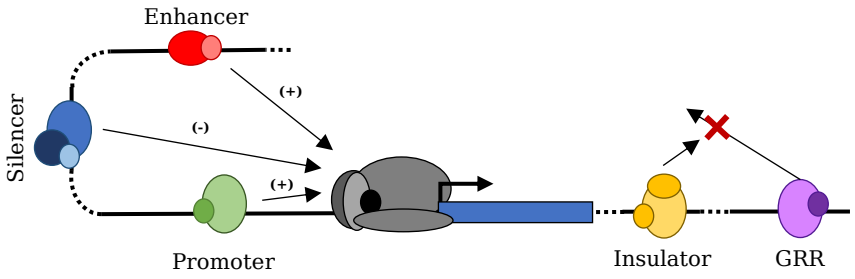


Figure 14: Overview of the transcriptional regulation by DNA binding proteins and associated cofactors. The final gene activity is defined by the combined effect of all regulators that bind to the regulatory region associated with that gene.

2.1.5.4 *Gene regulation via miRNAs*

The gene expression process can also be controlled by miRNAs. For this purpose, they form the so-called RNA-induced silencing complex (RISC) with different proteins, most notably members of the argonaute (AGO) family [555]. In this complex, the miRNAs guide the complex to a specific mRNA molecule that is then silenced. To this end, the miRNA binds to its target mRNA, mostly at the 3′ UTR. This is achieved via the complementary base pairing of the target mRNA and the seed sequence of the miRNA, i.e., bases 2-7 from the 5′ end of the mature miRNA. The RISC then mediates gene silencing in two different ways, either by translational inhibition or by mediating mRNA decay (cf. Figure 15). For both modes of actions, different mechanisms have been described [370]. In this manner, each miRNA can control a variety of target genes.

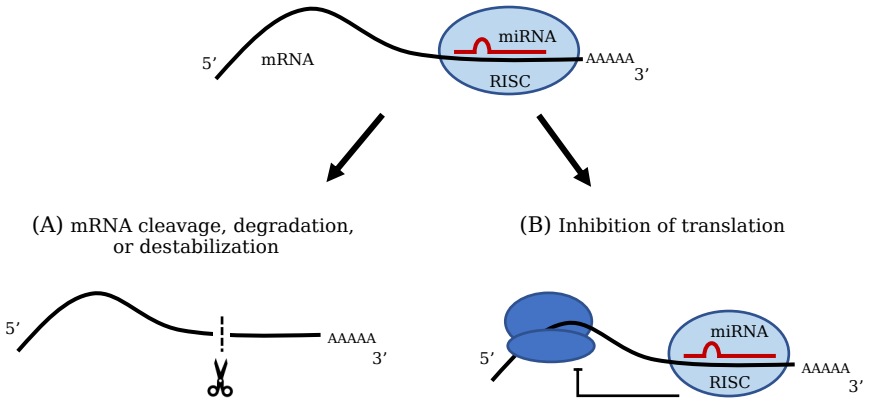


Figure 15: Regulatory effects of the RNA-induced silencing complex (RISC) formed by a miRNA and a member of the argonaute (AGO) family. (A) mRNA cleavage and (B) translation inhibition.

2.2 IMMUNE SYSTEM

The immune system is a complex network of molecular processes that combat infections, foreign particles, pathogens, or that destroy aberrant cells, e.g., cancer cells [116]. Most of these processes are governed and executed by different immune cells in close collaboration with specialized molecules. The majority of these cells are derived from hematopoietic stem cells in the bone marrow [373] (cf. Figure 16). However, some already develop during embryogenesis, in particular tissue-specific macrophage populations, like microglia [373].
Generally, the immune system is orchestrated in a layered hierarchy that, with each level, increases in specificity against pathogens [116]. The first layer of defense consists of (1) physical barriers that impede foreign particles, microorganisms, and viruses from entering the host organism, and (2) chemical compounds that combat pathogens, including antimicrobial peptides, like defensins [170]. The second layer is the innate or natural immune system that carries out an immediate but unspecific response. Cells and molecules belonging to this class always have a generic response to pathogens, independent of how often they are encountered. In contrast, cells of the adaptive immune system, which make up the third layer, can recognize and memorize specific pathogens. If these cells are presented with a known pathogen, they can initiate a strong and specific defense response. In the following paragraphs, we describe the function of molecules and different immune cells, i.e., leukocytes, in the innate or adaptive immune response.

Most information in the following paragraphs is based on the book 'Janeway Immunologie' by Kenneth Murphy and Casey Weaver [373] and supplemented with additional sources.

2.2.1 *Mechanisms of the innate immune system*

After pathogens have successfully breached the physical and chemical barriers of the first layer of defense, they are confronted with cells and associated molecules of the innate immune system. These leukocytes carry out various specialized functions to identify infections, alert other immune cells, and destroy pathogens or aberrant cells. In the subsequent paragraphs, we describe essential mechanisms of the innate immune system. Although most of the presented processes can be used to categorize the associated leukocytes, nearly all cells are directly or indirectly involved in multiple mechanisms.

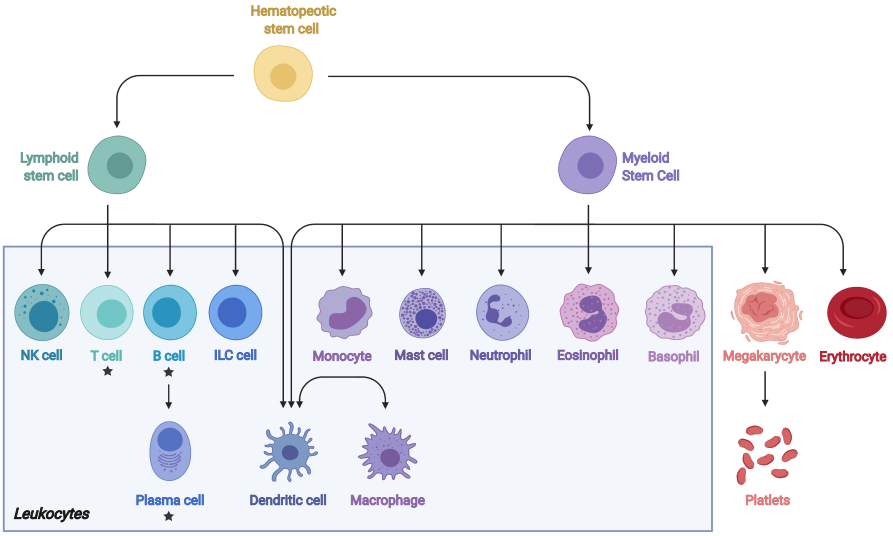


Figure 16: Overview of blood cells, including leukocytes (white blood cell), derived from hematopoietic stem cells. Cells of the adaptive immune system are marked with a black star. This figure was created using BioRender.com.

2.2.1.1 *The complement system*

The complement system is a central part of the innate immune system and plays a major role in inflammation and the defense against pathogens [242, 354]. It consists of a class of plasma proteins that circulate through the blood and other body fluids in an inactive state and act as a "surveillance system"[242, 354, 355]. Upon activation, they interact with each other to trigger a series of immune reactions that help to battle infection [242, 354, 355]. Amongst others, they promote inflammation, and they can destroy the membranes of microorganisms. Moreover, they cover and mark pathogens (opsonization), which facilitates their ingestion phagocytes [242, 354, 355].

2.2.1.2 *Phagocytosis and antigen presentation*

Phagocytes are a group of immune cells that monitor the bloodstream and tissues in the search for pathogens and diseased or damaged cells. Cells with primarily phagocytic activity are, for example, dendritic cells, neutrophils, and macrophages. Most phagocytes recognize foreign particles or cells via receptors on their surface that detect foreign particles. After phagocytes have recognized a foreign body, they can engulf and destroy them.
Some phagocytes, such as dendritic cells or macrophages, can also present small pieces of dismantled pathogens, i.e., epitopes, as antigens on their surface to alert or activate lymphocytes (cf. Section

2.2.2). Thereby, these antigen-presenting cells (APC) constitute a bridge between adaptive and innate immune processes.

2.2.1.3 *Promoting inflammation through chemical signals*

Many leukocytes can also produce and secrete chemicals that alert and attract other immune cells. These chemicals include cytokines that transmit signals between different leukocytes and chemokines that promote chemotaxis and, hence, facilitate the migration of cells towards the site of inflammation.

2.2.1.4 *Degranulation*

Certain types of leukocytes have granules, i.e., vesicles filled with enzymes or toxins, in their cytoplasm, for example, basophils, eosinophils, mast cells, or natural killer cells (NK cells). Upon activation, the granules are secreted, and the contained substances are dispersed. Depending on the substance, this can have different effects. For example, granules of NK cells, amongst others, contain antimicrobial agents like defensins and cytotoxins like perforin or granzymes. While perforin (PRF1) destroys the cell walls of neighboring cells, granzymes, e.g., GZMB, induce apoptosis in target cells.

2.2.2 *Mechanisms of the adaptive immune response*

Unlike innate immune cells, which have a general response independent of the encountered pathogens, adaptive immune cells, such as B and T lymphocytes, have a highly specific response to one particular antigen.
In general, an organism has many different adaptive immune cells that are either already specialized to respond to particular antigens or naive cells that have not yet encountered a pathogen.
After these naive lymphocytes are activated by a particular antigen, they go through a differentiation process, in which they divide into many daughter cells with specialized receptors that are able to recognize the respective antigen.
In this context, both types of lymphocytes can differentiate in either effector or memory cells. Effector cells can directly combat the pathogen but are short-lived, while memory cells last longer and prepare the organism for a future infection. In the following paragraphs, we will briefly discuss the different effector functions of B and T lymphocytes.

2.2.2.1 *Effector mechanisms of B cells*

Upon activation through their antigen receptor or T cells, B lymphocytes start to proliferate and develop into memory B cells or plasma cells, i.e., effector B cells (cf. Figure 17). The plasma cells can produce antigen-specific proteins, called antibodies. These antibodies bind to corresponding antigens and combat them through different mechanisms. Some antigens, such as toxins, can directly be neutralized by preventing them from affecting body cells. Additionally, antibodies can mark pathogens, like bacteria and viruses, for innate immune mechanisms such as phagocytes or the complement system.

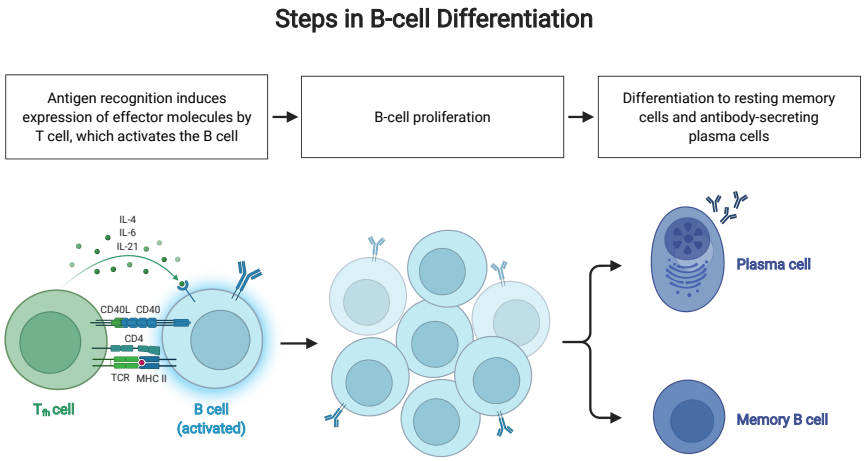


Figure 17: Overview of the B differentiation process. This figure was adapted from "Steps in B cell Differentiation", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates.

2.2.2.2 *Effector mechanisms of T cells*

T cells are activated by small peptides, called epitopes that are amongst others created by antigen-presenting cells (APCs), like phagocytes (cf. Section 2.2.1.2), and presented via MHC receptors on their surface (cf. Figure 18). In particular, there are two types of MHC receptors: (1) type I receptors, which are expressed in nearly all cells, and (2) type II receptors, which are mainly expressed in APCs. Different types of T cells recognize the two classes of MHC molecules. These lymphocytes express certain co-receptors (CD4 or CD8), which help their T cell receptor (TCR) bind the respective MHC molecules.
The co-receptors are also used to categorize the different T lymphocytes. CD8+ T lymphocytes are cytotoxic cells that recognize MHC type I molecules and kill all cells which present the respective epitope.

CD4+ T lymphocytes recognize epitopes presented by the MHC II receptor on APCs. Upon activation, CD4+ T cells produce cytokines that are secreted to alert or activate other leukocytes (cf. Figure 17) and to initiate or enhance their immune reactions.
CD4+ T cells can also be further divided into various subtypes that produce different sets of cytokines and fulfill different tasks (cf. Figure 18).

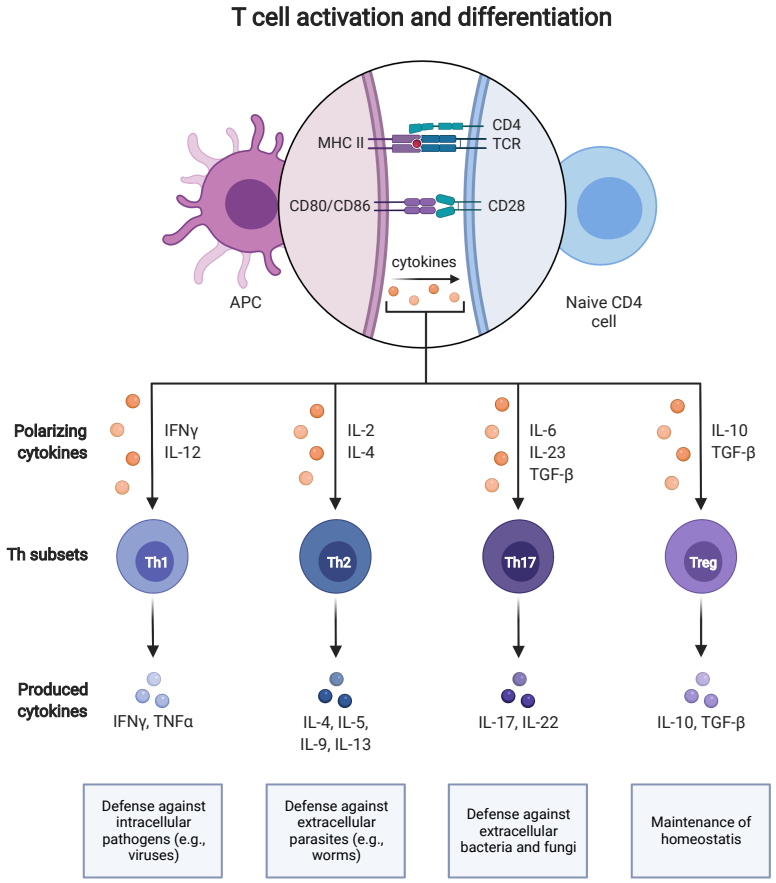


Figure 18: Overview of the CD4+ T cell activation and differentiation process. The functional characterization was obtained from [174] and the figure was adapted from "T cell activation and differentiation", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates.

2.3 CANCER

Cancer is a heterogeneous class of diseases that are characterized by abnormal cells with the ability to proliferate uncontrollably and the potential to invade other tissues. In the following sections, we briefly describe the mechanisms and characteristics of cancer cells that contribute to their development, progression, and malignancy.

Most information in the following sections is based on the book' The biology of cancer' by Robert A. Weinberg [569], the two landmark papers by Hanahan and Weinberg [201, 202], and supplemented with additional sources.

2.3.1 *Cancer development and hallmark properties*

The development of cancer cells is a continuous process in which healthy cells accumulate genetic and epigenetic alterations. In general, this is a natural process that automatically occurs over time. However, the emergence of new alterations can also be influenced by a manifold of external factors such as nutrition, infections, exposure to pathogens, or radiation [586]. In healthy cells, various defense mechanisms counteract this process and identify and repair these alterations [214, 492]. However, if these alterations are not corrected, they could cause disruptions in molecular mechanisms and ultimately lead to an acquisition of new traits that may alter the physiology, structure, or function of the affected cell and all descendants. While those new traits could be benign, they can also be responsible for cancer initiation or progression. Hanahan and Weinberg described several molecular characteristics, called the hallmarks of cancer, that are often acquired by cancer cells and contribute to their development and malignancy [201, 202]. In the following section, we provide a brief overview of these properties.

2.3.1.1 *The hallmarks of cancer*

In their landmark papers from 2000 [201] and 2011 [202] Hanahan and Weinberg describe eight hallmark characteristics of cancer cells and two enabling factors. An overview of these traits and examples for molecular factors that contribute to these capabilities are depicted in Figure 19 + 20.

SUSTAINING PROLIFERATIVE SIGNALING In normal cells, the proliferation is controlled via (external) signals that initiate or promote cell growth or division. These signals ensure that the number of cells and the structure of normal tissue is balanced. One property of many cancer cells is that they can sustain or even enhance this signal and thereby increase their proliferation activity. Examples of such alter-

ations are activating mutations of PI3KCA that transform this gene into an oncogene that promotes the proliferation of cancer cells [597].

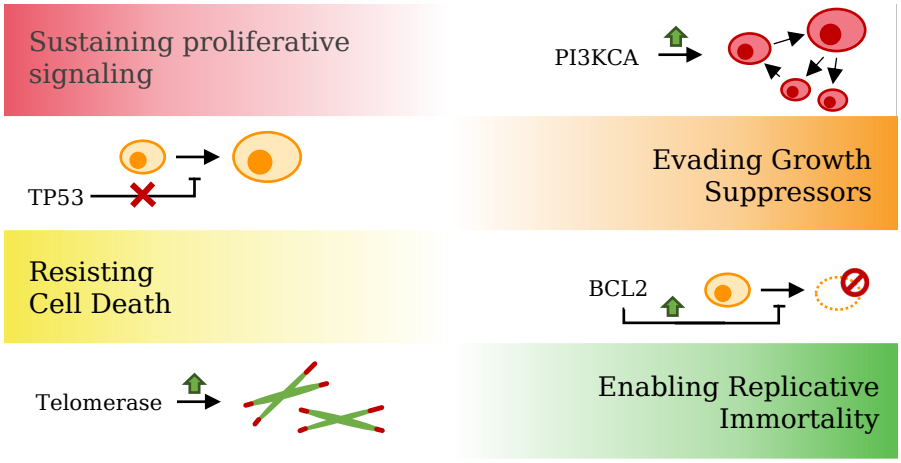


Figure 19: Examples of traits that contribute to the hallmark capabilities of cancer cells proposed by Hanahan and Weinberg [201, 202]

EVADING GROWTH SUPPRESSORS In general, the cell cycle is controlled by growth suppressors that can induce growth arrest or even apoptosis. These genes actively repress an uncontrolled proliferation of cells and inhibit the formation of cancerous cells. Due to this property, they are referred to as tumor suppressor genes (TSG). Consequently, in order to increase their rate of proliferation, cancer cells often exhibit loss-of-function mutations in TSG, which circumvents their regulatory role in this process. One prime example is the transcription factor TP53 that is regularly mutated in cancer cells [164, 194].

RESISTING CELL DEATH Normal cells, in general, have several mechanisms that prevent or counteract the development of cancer cells. A critical mechanism is the controlled induction of apoptosis, e.g., for cells with aberrant proliferation [315]. This process is regulated by an interplay of different transcription factors that balance the apoptotic and anti-apoptotic mechanisms in a cell. Cancer cells often exhibit disruptions of this regulatory circuitry and thereby prevent apoptosis. Examples for this are alterations that either increase the activity or expression of anti-apoptotic regulators such as BCL2 family members or suppress the activity or expression of apoptosis-inducing regulators, like BAK or BAX [4, 202, 315].

ENABLING REPLICATIVE IMMORTALITY Another property that many cancerous cells acquire is replicative immortality. This means that they can go through an unlimited amount of cell divisions without becoming senescent. To achieve this, cancer cells often show increased activity of the telomerase complex, which is typically not or

nearly not detectable in most mature cell types [267, 341, 584]. The telomerase adds repeat sequences to the end of chromosomes (telomeres), which protects them from erosion during cell division.

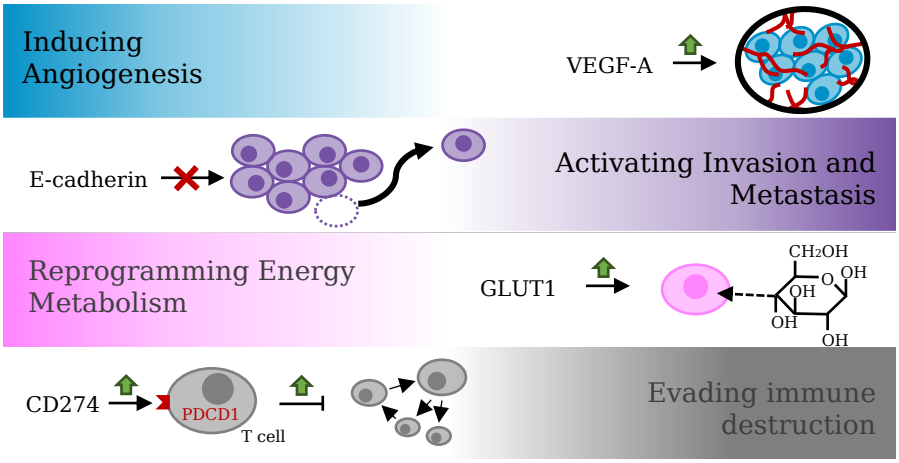


Figure 20: Examples of traits that contribute to the hallmark capabilities of cancer cells proposed by Hanahan and Weinberg [201, 202]

INDUCING ANGIOGENESIS To increase the rate of proliferation and to ensure their survival, cancer cells need an increased amount of oxygen and nutrients and an effective method to get rid of metabolic waste. For this purpose, cancer cells often induce angiogenesis, i.e., the extension of the vascular system to develop new blood vessels. This process is often facilitated by alterations that enhance the activity of associated activators such as VEGFA or, conversely, the repression of corresponding inhibitors like THBS1 [47].

ACTIVATING INVASION AND METASTASIS One property that heavily contributes to the malignancy of cancer is the invasion into other tissues and the formation of metastases. This is normally prevented by different protein families that help to tether cells to their surroundings [201]. Amongst those proteins are cell adhesion molecules, like CDH1 (E-cadherin), that are located on the cell surface and control the binding of a cell to its neighbors or the extracellular matrix. Hence, cancer cells often exhibit alterations in these molecules that suppress or inhibit their activity.

REPROGRAMMING ENERGY METABOLISM A further characteristic of cancer cells is that they often reprogram the energy metabolism. It has been shown that many cancer cells restrict their energy metabolism to glycolysis (Warburg effect), which in normal cells only occurs under anaerobic conditions [563]. While, on the one hand, this reduces the efficiency of ATP synthesis, it is hypothesized to facilitate the production of other macromolecules such as nucleosides and amino acids that are also required for the proliferation [544]. To compen-

sate for the energy metabolisms inefficiency, cancer cells often exhibit upregulation of the glucose transporter GLUT1, which increases the glucose uptake in the cell [202].

EVADING IMMUNE DESTRUCTION One crucial factor that counteracts cancer development is the immune system, which is in general able to recognize and destroy abnormal cells [477]. In order to evade this immune destruction, cancer cells often exhibit alterations that counteract this process, e.g., an increased expression of ligands for immune checkpoint proteins on the surface of immune cells [130]. If immune checkpoint proteins are activated, they negatively regulate the immune response, allowing cancer cells to evade immune destruction [477].

ENABLING FACTORS In addition to the eight hallmark characteristics, Hanahan and Weinberg also describe two enabling factors that help cancer cells to acquire the hallmark properties. The first one is genome instability, which facilitates the emergence of new alterations in the genome and thereby accelerates cancer development. The second enabling factor is tumor-promoting inflammation. While the immune system, on the one hand, combats cancer cells, it can also influence the tumor microenvironment and thereby promote the emergence of hallmark properties, e.g., by releasing growth factors that can boost the proliferation [202].

3

MATERIALS AND METHODS

The major goal of all tools presented in this thesis is the analysis of deregulated biological processes and the identification of driving factors within those processes. To this end, our tools process and analyze molecular high-throughput measurements of different omics[1] types, e.g., genomics, transcriptomics, or proteomics.
In this chapter, we first introduce the principles of several widely used experimental high-throughput techniques that can be applied for the molecular characterization of cells (cf. Section 3.1). Then, we describe the different data resources that can be used to annotate molecular measurements and facilitate their interpretation (cf. Section 3.2). Subsequently, we introduce some basics of hypothesis testing (cf. Section 3.3), and finally we provide a detailed description of different concepts for the computational analysis of molecular high-throughput profiles that are required for the remaining chapters of this thesis: (1) feature selection and group comparison (cf. Section 3.4 + 3.5), (2) enrichment analysis (cf. Section 3.6), (3) network analysis (cf. Section 3.7), and (4) regulator impact analysis (cf. Section 3.8). Additional methods can be found in Appendix B.

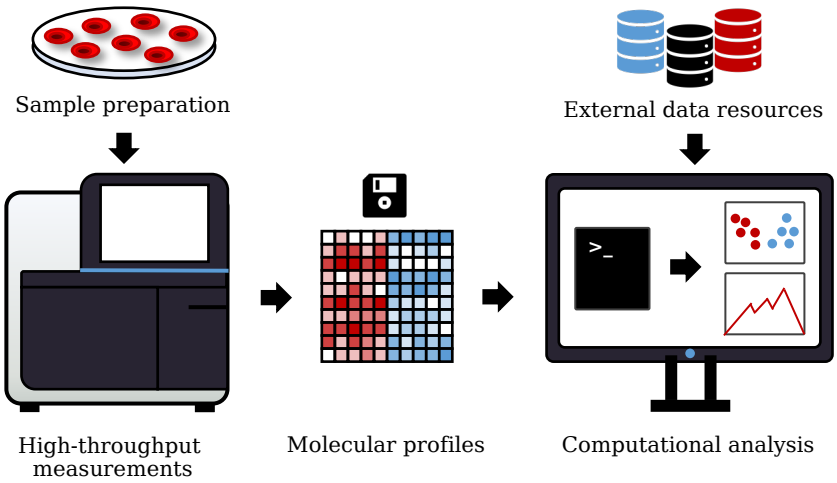


Figure 21: Overview of the Materials and Methods chapter. We first introduce the principles of several high-throughput assays that can be used to generate molecular high-throughput profiles. Then, we describe the external resources required for the analysis of these data sets. Finally, we discuss the computational approaches used throughout this thesis.

1 The term omics refers to different branches of biology that end in "omics", e.g., genomics, transcriptomics, or proteomics.

3.1 HIGH-THROUGHPUT ASSAYS

In this section, the principles of several high-throughput methods are presented that are used for the molecular characterization of cells. Here, we especially focus on technologies that were applied to generate the data sets analyzed throughout this thesis: high-throughput sequencing (cf. Subsection 3.1.1) and DNA microarrays (cf. Subsection 3.1.2). However, the functionality presented in the following chapters can also be applied to analyze measurements of other high-throughput platforms, such as mass spectrometry (see, e.g., [125, 179]).

3.1.1 *High-throughput sequencing*

High-throughput sequencing (HTS) assays are the most versatile and widely used of the high-throughput technologies. Data produced by these approaches builds the foundation of some of the biggest and most ambitious projects in current biological research, e.g., ENCODE [99], GTEx [314], TCGA [536], or The 100,000 Genomes Project [409]. Additionally, HTS assays are regularly applied in non-research settings, such as forensics [77] or even clinical applications [308].
There are currently a variety of widely used HTS platforms available, e.g., Illumina HiSeq [234], BGISEQ-500 [55], PacBio Sequel Systems [402], or Oxford Nanopore Technologies MinION [525].
In the following sections, first the principles of the Illumina sequencing by synthesis approach are introduced. Then different adaptations of the DNA sequencing protocol are presented that not only allow to sequence DNA, but also to measure epigenetic modifications or gene expression.

3.1.1.1 *Sequencing by synthesis*

The general workflow of this process is divided into four basic steps [93, 233, 436]: library preparation, cluster amplification, sequencing, and data analysis. In the following paragraphs, an overview of the different steps is provided.

Library preparation

In this step, the purified DNA is prepared for the subsequent sequencing procedure. First, the DNA is fragmented to generate small DNA pieces, which can then be sequenced in parallel. The fragmentation can either be accomplished by sonication [461] or tagmentation [93]. In both cases, sequencing adapters are ligated to both ends. Each adapter is a nucleotide oligomer (oligo) that consists of three distinct parts. The first part is used to fixate the DNA fragments on a glass slide, called the flow cell (cf. next section). The second part of the

adapter is a barcode that can be used to identify the sample the DNA fragment originated from. The barcode makes it possible to run multiple samples simultaneously. The third part is the binding site for the sequencing primer.

Cluster amplification

After the preparation step, the resulting DNA library is loaded onto a sequencing flow cell. The flow cell is a glass slide coated with oligos that are complementary to the two adapters at the end of each DNA piece. As the library is loaded onto the surface of this glass slide, the DNA fragments randomly hybridize to these oligos (cf. Figure 22 A). Next, an amplification step is conducted that creates a cluster of identical copies for each DNA fragment, i.e., clonal clusters. To this end, the adapter on the opposite side of each DNA strand is also hybridized to an oligo on the flow cell and builds a "bridge". Then a DNA polymerase is attached and synthesizes a complementary DNA strand (cf. Figure 22 B). Afterwards, the DNA double strands are denatured and the process is repeated until dense clusters of identical copies are formed. Finally, all reverse strands are removed, such that all DNA fragments in a cluster point to the same direction (cf. Figure 22 C).

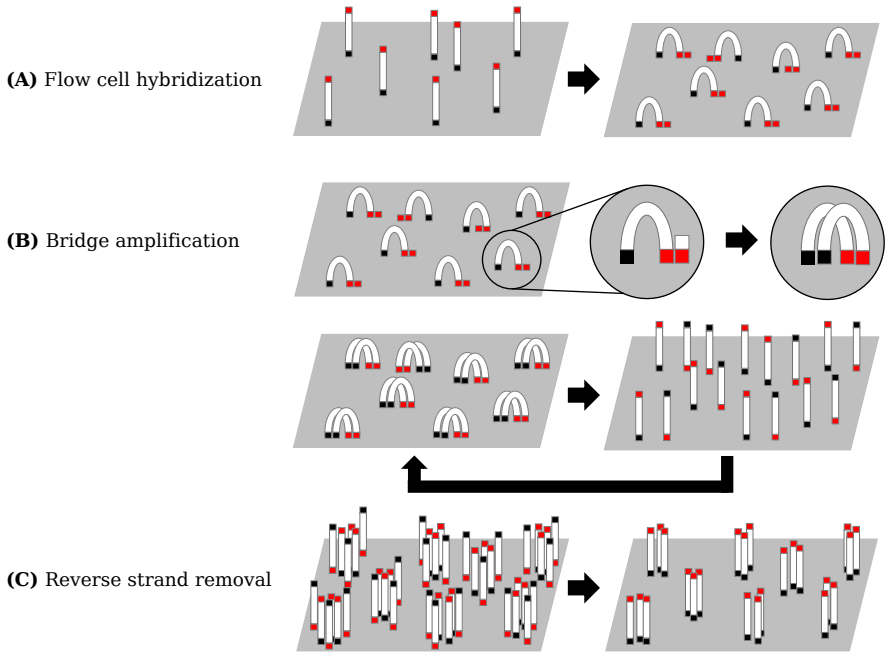


Figure 22: Cluster amplification steps of the Illumina sequencing pipeline. (A) Hybridization of the DNA fragments to the flow cell. (B) Amplification of DNA fragments (bridge amplification). (C) Removal of reverse strand to create clusters of unidirectional replicated DNA fragments.

Sequencing

The DNA fragments in each cluster are then processed simultaneously. In the sequencing step, for each DNA fragment a complementary strand is synthesized in a step-wise fashion. In each step, a fluorescently labeled nucleotide is incorporated. Each of the four nucleotides emits light with a specific wavelength that can be used to identify, which base was added. For each cluster, the incorporated nucleotide is then recorded to get the sequence of the DNA. This process is repeated until a read of a specific length is generated. An example of this workflow with five sequencing cycles is shown in Figure 23.

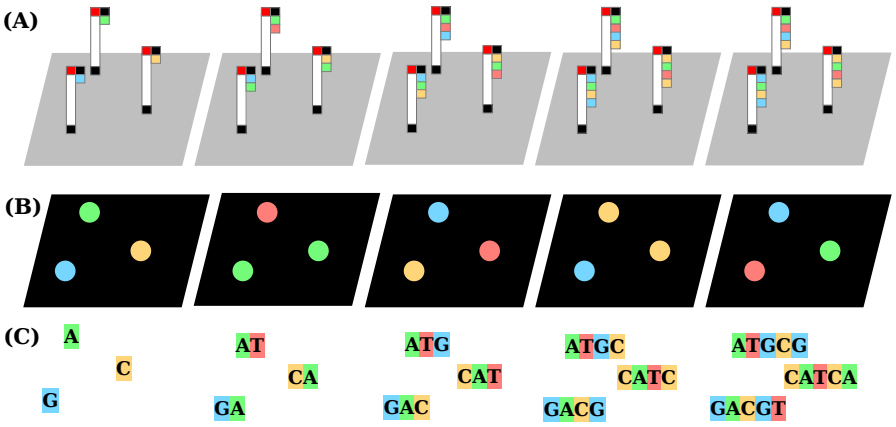


Figure 23: Example of five sequencing cycles of the illumina pipeline. (A) Step-wise integration of complementary nucleotides labelled with fluorescent die. (B) Image of the flow cell to record light emission. (C) Reads for each cluster.

The workflow outlined above describes the single-end sequencing protocol, where each DNA sequence is only sequenced from one end, the forward strand. However, for many applications the DNA fragments are sequenced from both sides, which is called paired-end sequencing. For this approach, additional steps are conducted. First, the newly synthesized DNA strand is removed and washed away. Then the forward strand, i.e., the original template strand, is used to synthesize its reverse complement analogously to the bridge amplification step. To this end, the DNA template (forward strand) folds over and binds the second type of oligo on the surface of the flow cell to form a bridge. In the next step, a DNA polymerase extends this oligo and synthesizes the reverse complement of the original template sequence. Finally, the forward strand is cleaved off and the sequencing steps are repeated for the reverse strand.

Data processing

One run of current Illumina sequencing platforms can produce up to 1 billion reads[2] with read length between 50 and 300 bp[3] [235]. After an initial filtering step that removes low quality reads, there are several options to process the remaining ones, including de novo assembly and alignment to a known reference genome.
In a de novo assembly, the genome is reconstructed based on the millions of generated short reads. To this end, overlapping reads are combined to continuous DNA sequences, called contigs. Based on additional information, e.g., provided by paired-end sequencing, contigs with matching orientation are then combined to larger scaffolds. For both tasks, various approaches have been proposed, e.g., EULER [410], MEGAHIT [297] or SOAPdenovo [302]. Most of them use the reads to generate a spectrum of k-mers that are used to build a de Bruijn graph [410]. This graph is then searched for Euler paths that can be used to construct contigs and scaffolds respectively [410]. Genome assembly approaches are often applied in microbiology, where the genome of the investigated bacteria or virus might not be known [276], or in metagenomics applications, where samples might contain genomes of multiple organisms [297].
In applications where a reference genome is available, the reads can also be aligned to this sequence. Most tools for this purpose build an index of the reference sequence that can then be used to find the best matches for each read. The most popular approaches use indices based on hash tables [218] or the Borrows-Wheeler Transformation (BWT) [289, 299]. Usually, after all reads are aligned to the reference sequence, a pileup[4] for each position in the reference genome is generated (cf. Figure 24). This data structure can then be used to analyze if and where the genome of the sequenced sample differs from the reference sequence. Using this technique, it is possible to deduce a variety of genomic aberrations, e.g., base substitutions, insertions, deletions, or even copy number variations (cf. Figure 24).

3.1.1.2 *Extensions, variations and modifications*

While DNA sequencing has been an invaluable tool in biology and medicine, there are several extensions, variations, and modifications of the previously described protocol that make these assays even more versatile and powerful. In the following paragraphs, we describe how standard HTS can be adapted to study not only the sequence of the DNA, but also epigenetic modifications of the DNA,

2 Maximum output of the Illumina NextSeq 2000 platform.
3 Maximum read length of the Illumina MiSeq platform.
4 A pileup is a data structure that maps sequence reads to the corresponding position of the genome.

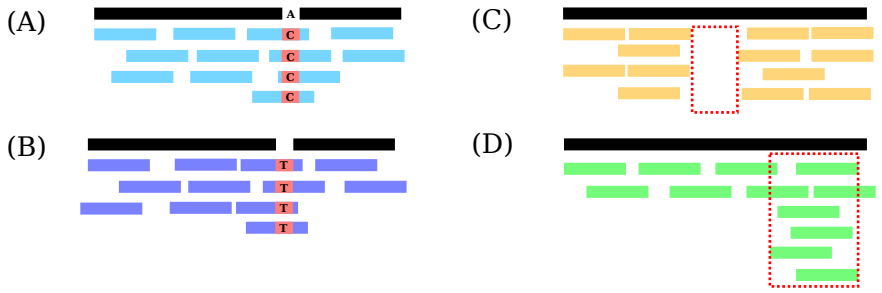


Figure 24: Overview of mutations that can be detected in a DNA sequencing experiment: (A) base substitution, (B) insertion, (C) deletion and (D) copy number alteration.

the location of DNA binding proteins on the genome, or even the amount of gene expression.

ChIP-seq

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) [405] is an experimental technique to assess and catalog DNA binding sites of specific proteins, such as a particular transcription factor or even a specific histone variant. To this end, a variety of processing steps have to be conducted. First, the sample is treated with formaldehyde to cross-link the DNA and all associated proteins in order to preserve the structure of DNA-protein interactions. In the second step, the DNA is fragmented to obtain DNA sequences that are suitable for the subsequent sequencing procedure. In the third step, the protein of interest is marked with a specific anti-body. Marked DNA-protein complexes can then be isolated (or purified). This process is called chromatin immunoprecipitation. The purified DNA-protein complexes are then heated to remove the cross-linking and to separate proteins and DNA. In the next step, extracted DNA fragments are sequenced and the resulting reads are aligned to the reference genome. Finally, the pileup of the mapped read is analyzed to identify genomic regions that show a significant enrichment, i.e., peaks. The peaks correspond to binding sites of the investigated protein. An overview of the protocol is shown in Figure 25.
For the identification of peaks a variety of methods have been developed [534, 542, 611]. All of them use statistical models to identify genomic regions that have significantly more mapped reads than control samples or a specific background window.

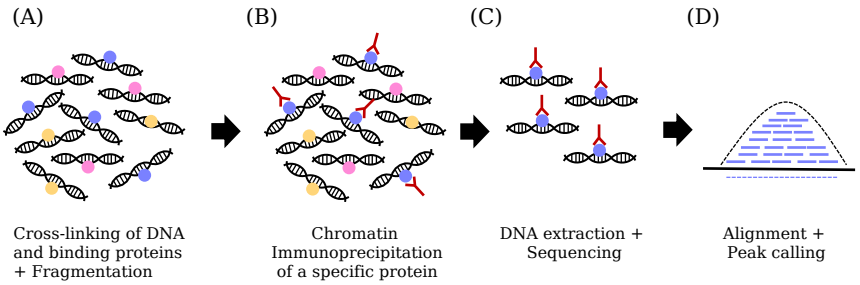


Figure 25: Overview of the ChIP-seq protocol. (A) DNA and associated proteins are cross-linked and the DNA is fragmented. (B) Chromatin immunoprecipitation is used to isolate DNA fragments that bind to the investigated protein. (C) DNA and proteins are then separated and the extracted DNA is sequenced. (D) Genomic regions that show a significant enrichment of aligned reads (peaks) correspond to binding sites of the investigated protein.

DNase-seq

The DNase-seq protocol can be used to analyze open or accessible regions of the chromatin [500]. To this end, the sample is treated with DNase I, an endonuclease that is able to cleave the chromatin at regions that are open (cf. Section 2.1.2) and not occupied by any DNA binding proteins. In the next step, the DNA is extracted and sequenced. The pileup of aligned reads can then be analyzed to identify regions that are significantly enriched with DNase I cut sites (DNase I hypersensitive sites). Accordingly, these peaks correspond to accessible or open chromatin regions. An overview of the protocol is depicted in Figure 26.
However, the read coverage of a DNase I hypersensitive site allows not only to analyze if the respective region is open, but also to identify so called footprints of DNA binding proteins [75, 131]. Footprints are narrow regions that are partially protected from DNA cleavage. They can be detected as indentations in the coverage of DNA (cf. Figure 26C). Several tools for this purpose have been proposed [196]. While footprints have successfully been used to identify the corresponding DNA-binding proteins [71], it has to be noted that not all proteins leave footprints when interacting with DNA [516]. Sung et al. [516, 517] successfully showed that the footprint depth is correlated with the residence time[5] of these proteins. It has also been reported that for dynamic proteins, i.e., proteins with small residence times, no footprints could be detected [209, 517].

5 The residence time indicates how long a protein is usually connected to the DNA.

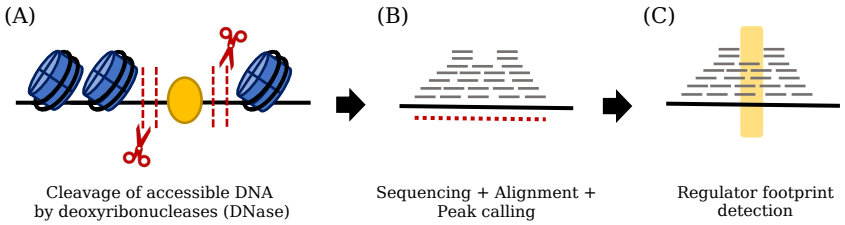


Figure 26: Overview of the DNase-seq protocol. (A) The chromatin is cleaved in regions that are accessible to the DNase I endonuclease. (B) The resulting DNA fragments can then be sequenced and aligned to the reference genome. Identified peaks correspond to open chromatin regions. (C) The pileup of aligned reads can also be analyzed to identify footprints of DNA binding proteins.

RNA-seq

In a RNA-seq experiment, the RNA contained in the cells is sequenced instead of the DNA [562]. This not only allows to determine the nucleotide sequence, but also to quantify the amount of different RNA molecules, i.e., the gene expression. To this end, the library construction protocol has to be adapted. Depending on the use case, either all RNA molecules or a specific subset (miRNAs, piRNAs, polyA, etc.) are extracted from the analyzed cells [562]. The extracted RNAs are then converted into cDNA. While small RNAs (e.g., miRNA) can directly be sequenced, larger RNAs need to be fragmented [562] to account for the used sequencing platform. This can either be done directly on the RNA level (e.g., using RNA hydrolysis or nebulization) or on the corresponding cDNA (e.g., using DNase I treatment or sonication) [562]. The resulting cDNA fragments are then sequenced using any of the modern high-throughput sequencing platforms. Finally, the resulting reads are aligned to the reference genome and subsequently assigned to genes in order to obtain expression estimates.
Common scenarios for the analysis of RNA-seq data are the comparison of expression values between different sample groups, e.g., disease vs. control, or between genes within the same sample. In both cases, expression values need to be normalized in order to avoid biases introduced during library preparation. Since samples may vary in the RNA content or sequencing depth, all expression values need to be scaled with respect to the library size in order to make them comparable. For the comparison between genes in the same sample, expression values need to be adjusted according to their length.[6] This is especially important for the analysis of mRNA, where the fragmentation step might result in more sequencing reads for longer transcripts. An example is depicted in Figure 27.

6 The length normalization can be avoided, when the fragmentation is omitted, e.g., for small RNAs or adapted RNA-seq protocols that only sequence the 3′ or 5′ ends of transcripts [364]

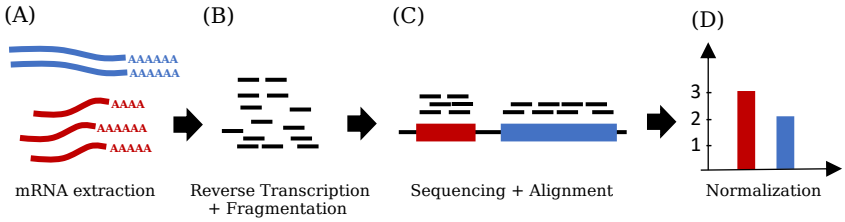


Figure 27: Example of the RNA-seq workflow for mRNAs. (A+B) Here, two transcripts with different lengths are sequenced. (C+D) Since more reads are produced from the longer transcripts, they need to be adjusted accordingly.

One method that is often used to account for both biases is the transcript per million (TPM) measure proposed by Wagner et al. [554]. For a set of genes (G), it can be calculated in a two-step approach. First, for each gene ($g \in G$), the number of mapped reads per gene ($R(g)$) is adjusted by the respective gene length ($l(g)$) in kilobases:

$$E_l(g) = \frac{R(g) * 1000}{l(g)}$$

The length normalized expression values $E_l(g)$ are then used to scale gene values with respect to the library size in each sample:

$$TPM(g) = \frac{E_l(g)}{\sum_{\tilde{g} \in G} E_l(\tilde{g})} \times 10^6$$

While the 'per million' scaling factor is appropriate for library sizes in standard bulk RNA-seq data sets, it is sometimes adapted for alternative assays. For single cell RNA-seq experiments, a smaller scaling factor might be more appropriate. In this context, the expression values of single cells are often scaled to the median library size of all cells in the data set [320]. After the normalization, $TPM(g)$ values are usually $\log_2$ transformed:

$$\log TPM(g) = \log_2(TPM(g) + 1)$$

Here, the 1 is added as a pseudo-count to avoid problems with 0 values. There are also a variety of approaches that can be used instead of TPMs, such as TMM [444] or GeTMM [495]. More information on these methods can be found in the respective publications.

In order to avoid the fragmentation bias concerning the transcript length, alternative experimental protocols have been proposed that omit the fragmentation step and instead only sequence the 3′-end [322, 357, 364] or 5′-end [5, 372] of each transcript. In particular 5′-end RNA-sequencing methods have the additional advantage that they can help to identify the correct transcription start site of each transcript. However, both approaches only provide limited sequence

information. Hence, depending on the use case it is crucial to select the correct protocol.

3.1.1.3 *Single-cell sequencing*

Conventional bulk-sequencing experiments, as described in the previous sections, measure the average signal in a population of cells. However, a sample might contain sub-populations of cells with a completely different phenome, genome, transcriptome, or epigenome. This means that information could potentially be lost or that a signal might be distorted in a certain direction depending on the cell composition. For gene expression measurements it has been shown that a few cells can have a strong influence on the mean expression level in a population [43, 585]. Variances between cells have even been detected for immortalized cell lines that are used as gold standard for somatic mutation calling or drug screening (e.g., COLO − 829 [548]). To this end, single cell sequencing approaches have been developed. These allow to analyze the signals of individual cells, to detect different sub-populations such as cell types and to investigate variations between the cells in a sample.

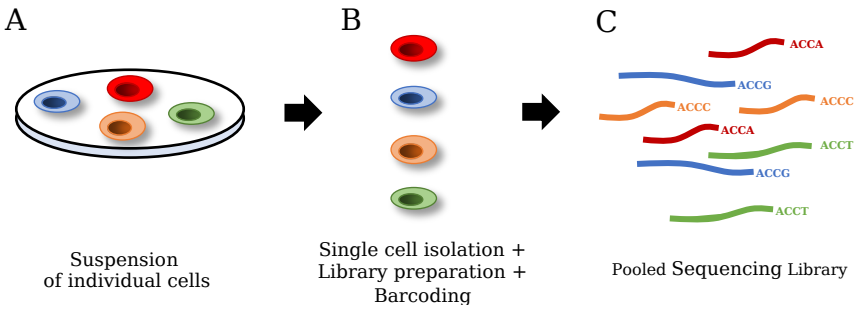


Figure 28: Overview of single cell isolation and library preparation step. (A) The protocol starts with a suspension of individual cells. (B) Each single cell is then isolated, the target DNA is extracted and barcoded. (C) Finally, the target DNA of all cells is pooled to create the single cell sequencing library.

Workflow Overview

The first step in most single cell protocols is the preparation and isolation of single cells. To this end, a variety of techniques have been proposed: flow-activated cell sorting (FACS) [65], laser capture microdissection (LCM) [140], microfluidics- [577] or split-pool approaches[401]. The choice depends on the experiment design and the amount of cells that should be sequenced [224]. After the cells are isolated, modified versions of the library preparation protocols described in the previous paragraphs can be applied: scRNA-seq [460], scDNAse-seq[245], scATAC-seq [81], or scChIP-seq [450]. One important modification in nearly all single-cell protocols is that the resulting cDNA is barcoded

such that the originating cell can be identified (cf. Figure 28). The barcoded DNA of all cells is then combined and sequenced as described before (cf. Section 3.1.1).

Dropout rate

While single-cell technologies provide a previously unmet resolution of the amount of biomolecules (DNA, RNA, protein, ...) in a cell, they also create various new computational challenges that make the analysis of single-cell data more complex. In particular, this is caused by the high sparsity of the generated data sets. Depending on the used technology and throughput of experiments overall dropout rates[7] between 82% and 97.41% have been reported for scRNA-seq experiments [425].

Single-cell multi-omics protocols

Recent advancements make it possible to measure multiple omics types simultaneously (cf. Table 1). This creates new possibilities for research in biology and medicine at a resolution that was formerly not possible [62]. These approaches bare the potential to uncover new links between the different omics types and even promise to reveal new insights into complex diseases, such as cancer [62]. For this reason, the editors of the journal "Nature Methods" have elected single-cell multimodal omics technologies as the method of the year 2019 [527].

	DNA sequence	Copy number variations	Chromatin accessibility	Chromosome conformation	Methylation	Gene expression	Surface proteins
10x Multiome	i*		x			x	
CITE-seq [511]	i*					x	x
Methyl-HiC [298]	i*			x	x		
scNMT-seq [95]	i*		x		x	x	
scTrio-seq [220]	x	i			x	x	

Table 1: Overview of single-cell multimodal omics protocols. Cells marked with 'x' specify omics types that were measured directly. Cells marked with 'i'/'i*' indicate omics types that are not directly measured, but can be inferred genome wide / partially.

7 The dropout rate defines the number of zeros in a single-cell experiment.

3.1.2 *Microarrays*

Next to modern sequencing techniques, further very popular high-throughput assays are so called microarrays or biochips. In general, microarrays are slides made out of glass, polymer, or nitrocellulose that are coated with thousands of specific DNA or protein probes [39]. These probes can be used to measure the amount of hybridized complementary DNA in case of DNA microarrays or the amount of bound proteins in case of protein arrays. Consequently, a variety of microarrays have been developed that can be used to measure different omics types, e.g., epignomics [236], genomics [7], transcriptomics [9], or proteomics [433, 528].
In this section, we focus on DNA microarrays [470] that are used to measure the expression of miRNAs and mRNAs. In both cases the probes are small DNA fragments that are highly specific for certain genes or even transcripts. Multiple copies of each probe are attached on the slide and form spots. Usually, each gene (or mature miRNA) is represented by several spots with different probes that are distributed over the glass slide [282].
The first step in a microarray experiment is either the isolation of all RNA molecules (total RNA) or a specific group of RNAs (e.g., mRNA or miRNA) in the analyzed sample. The extracted RNA is then converted into cDNA using a reverse transcriptase, labelled using a fluorescent dye (e.g., Cy3 or Cy5), and amplified. In the next step, the cDNA molecules are loaded onto the microarray slide, where they hybridize to complementary probes (Figure 29). After the hybridization, unbound molecules are washed away. In the last experimental step, each spot is scanned with a laser and the intensity of the light signal is recorded.

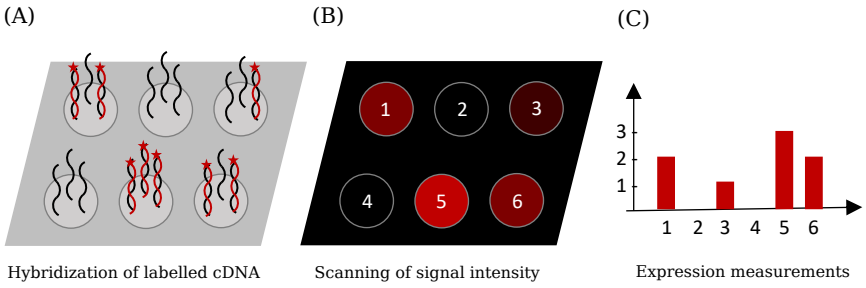


Figure 29: Overview of the Microarray workflow. (A) Hybridization of fluorescent labelled cDNAs (red). (B) Scanning of emitted signal intensity. (C) Extracted gene expression measurements.

The raw expression values need to be normalized in order to ensure comparability between different probes and different arrays. For the former, a background correction is performed that removes the effects of unhybridized cDNA molecules that are not properly washed away. For the latter, different techniques have been developed that

transform the expression values of each considered array into the same range. Two very popular approaches for this purpose are quantile normalization as for example implemented in the limma package [498] or variance stabilizing normalization [229].
Similar to RNA-seq experiments, normalized expression values are usually $\log_2$-transformed. Hoyle et al. [221] have shown that the distribution of microarray expression values can be approximated with a log2-normal distribution although the tail is more similar to a power law distribution. This means that logarithmized expression values are approximately normally distributed and statistical methods that assume this distribution can be applied.

3.2 THIRD-PARTY RESOURCES

Many tools for the analysis of molecular high-throughput profiles, including the ones described in this thesis, rely on prior knowledge of the measured features to improve or augment the analysis, or to facilitate the interpretation of computational results.
In this section, we describe data types and databases we have utilized to build our web services.

3.2.1 *Identifier*

Each biological feature, e.g., gene, protein, or miRNA, measured in a high-throughput experiment is typically characterized by a unique identifier. Depending on the used platform, manufacturer, or database, different IDs are used to denote a specific molecular entity.
Hence, one of the first steps in all our workflows is the validation and standardization of the used identifier types. To this end, we extracted reference lists and mapping tables from several databases: NCBI Gene [326], UniProtKB [68], Ensembl [12], miRBase [281], miRCarta [28], and miRTarBase [228]. These are saved in an in-memory database that can be utilized to map between the different annotations. All supported identifier types are listed in Appendix G.2 and our mapping procedure is described in Section 4.3.2.

3.2.2 *Reference sets*

Depending on the used protocol or platform, high-throughput experiments can measure a different number of molecular features. For example, RNA-seq experiments are often restricted to a particular subset of RNAs (miRNA, non-coding RNA or poly-A; cf. Section 3.1.1.2) and microarrays per definition only measure a preselected subset that is specified by the manufacturer.
Hence, we collected a variety of reference sets that compile all features measured for a given experimental protocol. These include mi-

croarrays from manufacturers like Affymetrix and Agilent that were extracted using the Ensembl BioMart tool [269], as well as genomic regions for genes and associated regulatory regions from Gencode [207], RefSeq [421], and Genehancer [154].

3.2.3 *Biological processes, signaling pathways, and functional categories*

An important type of information for each molecular feature is if it contributes to particular biological processes or signaling pathways. For our framework, we collected a variety of databases that assign molecular features to a particular biological function, e.g., signaling pathways from KEGG [394], Wikipathways [256], or Reactome [142], and biological processes, cellular components, and molecular functions from the Gene Ontology [100]. A complete list of the integrated resources can be found in Appendix G.3.

Most databases define biological processes or molecular functions as sets of features, we call categories (cf. Figure 30A). Other databases, in particular KEGG, also provide interactions of features as a directed network, where vertices represent features and edges specific interactions among them (cf. Figure 30B).

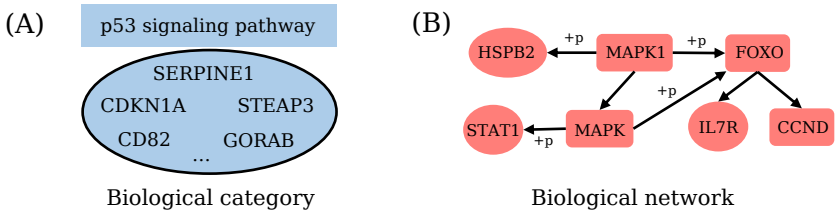


Figure 30: Examples of biological categories and regulatory networks. (A) Feature subset of the KEGG p53 signaling pathway and (B) subgraph of the KEGG regulatory network.

3.2.4 *Binding sites of transcriptional regulators*

In order to analyze the effects of transcription factors, chromatin modifiers, and co-factors we rely on different types of binding information discussed in the following paragraphs.

3.2.4.1 *Regulator-target interactions (RTIs)*

The first type of binding sites are regulator-target interactions (RTIs). We define a RTI as an experimentally determined binding site of a specific regulator within a regulatory region of a gene (cf. Figure 31).

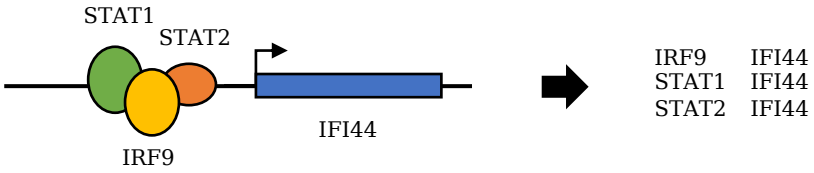


Figure 31: Examples of regulator-target interactions (RTIs).

For our framework, we collected binding sites from several databases: ChEA [286], ChipAtlas [396], ChipBase [592], ENCODE [99], JASPAR [158], Signalink [146], and TRANSFAC [342]. These databases either contain manually curated RTIs from literature or binding sites extracted from ChIP-seq experiments. Depending on the database, the binding sites are either reported as raw peaks or directly as RTIs. For raw peaks, we utilize BEDTools [426] to determine if any binding site of a regulator overlaps with the promoter region of a gene. For this purpose, the promoter of a gene is defined as a window around the transcription start site (TSS). For the RegulatorTrail web service, we use the following windows: TSS -/+ 1000, TSS -/+ 5000, TSS -/+ 10000, TSS -10000 / +1000. An overview of the information provided by the individual databases is shown in Appendix G.4.

3.2.4.2 *Transcription factor motifs*

Besides experimentally validated binding sites, we also use sequence motifs to study the binding patterns of regulators. A motif constitutes enriched sequence patterns at binding sites of a regulator (cf. Figure 32 A). Amongst others, they can be extracted from sequences of ChIP-seq peaks of the respective regulator [36].
Sequence motifs are often represented as position count matrices (PCM) or position weight matrices (PWM) that for each position of the motif indicate the frequency of the different bases (cf. Figure 32 B). Binding motifs can be utilized to predict binding sites of a regulator. This is for example useful for cell types where no experimentally validated binding sites are available. Different approaches for this purpose are discussed in Section 3.8.2.
For RegulatorTrail (cf. Chapter 8), we collected PWMs from several data sources: HOCOMOCO [283], JASPAR [158], Kellis Lab ENCODE Motif Database [265], and TRANSFAC [342].

(A)

(B)

A	14	4	3	56	56	3
C	21	2	0	1	0	18
G	19	48	52	0	0	34
T	3	3	2	0	1	2

Figure 32: Binding motif of the transcription factor SPI1 extracted from the JASPAR database [158]. (A) Sequence logo and (B) Position count matrix.

3.2.5 *miRNA-target interactions (MTIs)*

Next to transcriptional regulators, we are also interested in the analysis of miRNAs and their target genes. For this purpose, we collected two different types of miRNA-target interactions (MTIs): (1) Experimentally validated miRNA targets from miRTarBase [228] and (2) predicted miRNA targets generated using MiRanda [53] and TargetScan [8].
The different MTI sets build the foundation of the miRPathDB database (cf. Chapter 6). The experimentally validated targets sets are also used in the GeneTrail web service (cf. Chapter 5).

3.2.6 *Drug- and disease-related resources*

For DrugTargetInspector [474] and ClinOmicsTrail [475] (cf. Chapter C), we gathered drug- and disease-related annotations of molecular features, especially in the context of cancer. These include cancer driver genes from IntOGen [185], functional annotation for genetic variants from COSMIC [157] or dbSNP [484], molecular drug targets from DrugBank [582], molecular features that influence the efficacy of specific drugs from PharmGKB [211] or GDSC [238]. Additionally, our tools rely on clinical practice guidelines from the American Cancer Society (ACS) [3] and European Society for Medical Oncology (ESMO) [348, 482].

3.3 FUNDAMENTALS OF HYPOTHESIS TESTING

In this section, we introduce several mathematical concepts that are required for the subsquent sections. In particular, we introduce the concept of hypothesis testing and we discuss the multiple testing problem.

3.3.1 *Hypothesis tests*

A hypothesis test is a statistical inference method that tests an assumption about a population parameter based on a random sample [386, 446]. The following formal definition is based on the fourth edition of the book "Probability and Statistics" by Morris DeGroot and Mark Schervish [112].
Given a statistical problem involving an unknown parameter θ that is part of a parameter space Ω. Additionally, suppose Ω can be partitioned into two disjoint subsets Ω_0 and $\Omega_1 = \Omega \setminus \Omega_0$. Based on this information, we can define two hypotheses:

$$H_0 : \theta \in \Omega_0 \text{ (null hypothesis)} \tag{1}$$

and

$$H_1 : \theta \in \Omega_1 \text{ (alternative hypothesis)} \tag{2}$$

Since Ω_0 and Ω_1 are disjoint, θ can only be part of one of the two sets and either H_0 or H_1 must be true. In the hypothesis testing problem, we now try to decide if θ is part of Ω_0 or Ω_1. When we decide that $\theta \in \Omega_1$, we reject the null hypothesis. Otherwise, we do not reject H_0 (cf. Figure 33).

Depending on the task for which a hypothesis test should be applied, the null and the alternative hypothesis can be formulated one-sided or two-sided. In a one-sided case, the assumptions of both hypotheses include directions:

$$H_0 : \theta \leqslant \theta_0 \quad \text{and} \quad H_1 : \theta > \theta_0 \tag{3}$$

$$H_0 : \theta \geqslant \theta_0 \quad \text{and} \quad H_1 : \theta < \theta_0 \tag{4}$$

Accordingly, the hypotheses of two-sided tests are of the form:

$$H_0 : \theta = \theta_0 \quad \text{and} \quad H_1 : \theta \neq \theta_0 \tag{5}$$

In a hypothesis test, we base the decision if H_0 should be rejected on information we obtain from a random sample $X = \{x_1, ..., x_n\}$ drawn from a distribution that involves θ. To this end, we use a testing procedure δ that divides the sample space S of X into two disjoint sets $S_1 \subset S$ and its complement $S_0 = S \setminus S_1$, such that if $X \in S_1$, we reject the null hypothesis. We refer to S_1 as the "critical region".

This critical region is often defined in terms of a test statistic $T : S \to \mathbb{R}$, i.e., a function that assigns a real value to each possible sample in S. Additionally, let $R \subset \mathbb{R}$ be the set of all possible values of $T(X)$ for which we reject H_0, which we call the rejection region. Based on R, we can now define S_1 as:

$$S_1 = \{x : T(X) \in R\} \tag{6}$$

In practice, the rejection region is a fixed interval in the value range of the test statistic, e.g., we might reject H_0 for all values of $T(x)$ that are larger or equal than a constant c ($T(X) \geqslant c$). In this case, the interval $[c, \infty)$ would be our rejection region.

Given a test procedure δ_c with test statistic $T(x)$ that rejects H_0 if $T(X) \geqslant c$, we can define a power-function $\pi(\theta, \delta_c)$ that specifies the probability that δ will reject H_0 for a given $\theta \in \Omega$:

$$\begin{aligned} \pi(\theta, \delta_c) &= P(X \in S_1|\theta) && (7)\\ &= P(T(X) \in R|\theta) && (8)\\ &= P(T(X) \geqslant c|\theta) && (9) \end{aligned}$$

Accordingly, $1 - \pi(\theta, \delta_c)$ gives us the probability that δ_c does not reject the null hypothesis.
By using any test δ_c, we can make two kinds of errors: (1) a type I error when δ_c wrongfully rejects H_0, and (2) a type II error when δ_c does not reject a false H_0. The probabilities for these errors are denoted with $\alpha(\delta_c)$ and $\beta(\delta_c)$ respectively:

$$\begin{aligned} \alpha(\delta_c) &= \sup_{\theta \in \Omega_0}(\pi(\theta, \delta_c)) && (10)\\ \beta(\delta_c) &= \sup_{\theta \in \Omega_1}(1 - \pi(\theta, \delta_c)) && (11) \end{aligned}$$

Suppose, we have the choice between multiple tests, then we need to select a test δ_c that balances between $\alpha(\delta_c)$ and $\beta(\delta_c)$. To this end, $\alpha(\delta_c)$ is typically set to a fixed upper-bound α_0, which is called the significance level. For the testing problem this means that we try to find the best test procedure δ_c with the following constraint:

$$\sup_{\theta \in \Omega_0}(\pi(\theta, \delta_c)) \leqslant \alpha_0 \tag{12}$$

Now, given a test procedure δ_c with test statistic $T(X)$, a random sample $X = \{x_1, ..., x_n\}$ drawn from a distribution that involves θ, and a significance level α_0. In order to test if we can reject H_0 based on the test value $t = T(X)$, we need to compare t and α_0. To this end, we use a measure called p-value that is defined as the probability to obtain a test value at least as extreme as t, assuming H_0 is true:

$$p = \sup_{\theta \in \Omega_0}(\pi(\theta, \delta_t)) \tag{13}$$

If the obtained value is smaller than the significance level α_0, we reject H_0.

In accordance with the H_0 and H_1, the p-value can be calculated one-sided or two-sided. For a one-sided, upper-tailed test with test statistic $t = T(X)$, the p-value is defined as:

$$p = sup_{\theta \in \Omega_0}(P(T \geqslant t|\theta)) \tag{14}$$

(A)

	H_0 is true	H_0 is false
H_0 not rejected	$1 - \alpha$	β
H_0 rejected	α	$1 - \beta$

(B)

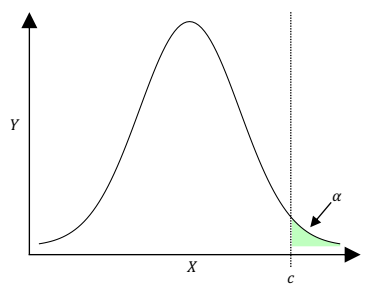


Figure 33: (A) Table summarizing the decisions that can be made in a hypothesis test. (B) Illustration of a critical region in a hypothesis test.

3.3.2 *Multiple testing correction*

In practical applications, such as the analysis of high-throughput experiments discussed in this thesis, multiple hypothesis tests are often conducted simultaneously, e.g., to infer which biological pathways are significantly deregulated in an analyzed data set (cf. Section 3.6 or Chapter 5).
From the last section, we know that for each test δ_c a predefined significance level (α_0) is used to judge if a result significantly deviates from the null hypothesis. This means that in each test, we also have the probability α_0 to get a false positive result (type I error). For k conducted tests, the probability accumulates and we get the following probability to make at least one false positive decision, which is called the family-wise error rate (FWER):

$$p_{FWER} = 1 - (1 - \alpha)^k \tag{15}$$

Over the years, different methods have been proposed to control the FWER [152, 217, 487]. Here, one of the most popular methods is the Bonferroni correction [63, 64]. Assuming we have conducted n statistical tests with the p-values $p_1, p_2, ..., p_n$. Then each p-value p_i is adjusted with the number of tested hypotheses:

$$\tilde{p}_i = p_i * n \tag{16}$$

The null hypothesis of a test is only rejected if ($\tilde{p} < \alpha$). Hence, another common usage of this method is that the significance level is adjusted, i.e., we test if $p < \frac{\alpha}{n}$.

Another popular approach to reduce the number of false decisions in a multiple testing scenario is to control the false discovery rate (FDR) [44]. It is defined as the expected proportion of discoveries that are falsely rejected [44, 278]. The most popular approach to control the FDR is the correction method proposed by Benjamini and Hochberg [44]. In contrast to the Bonferroni correction, this method adjusts each p-value with a specific correction factor rather than processing all in the same manner. Assuming we have n independent p-values that are sorted increasingly, then the p-values are adjusted in descending order as follows:

$$q_i = \begin{cases} p_i & \text{if } i = n \\ \min\{q_{i+1}, \frac{n}{i} p_i\} & \forall\, i \in \{n-1, ..., 1\} \end{cases} \tag{17}$$

The individual q_i can be interpreted as the expected FDR for tests with this value. Hence, the null hypothesis is rejected if $q_i < \alpha$, where α in this case is the accepted FDR.
The Benjamini-Hochberg method is proven to have a much higher statistical power than many methods controlling the FWER, but only for independent tests [45]. However, the method can be adapted, such that it can also be applied if the conducted tests are statistically dependent. To this end, Benjamini and Yekutieli proposed an extension of this approach that uses an additional correction factor γ to adjust p-values [45].

$$\gamma = \sum_{i=1}^{n} \frac{1}{i} \tag{18}$$

$$q_i = \begin{cases} \gamma p_i & \text{if } i = n \\ \min\{q_{i+1}, \gamma \frac{n}{i} p_i\} & \forall\, i \in \{n-1, ..., 1\} \end{cases} \tag{19}$$

3.4 STATISTICAL FEATURE SELECTION AND GROUP COMPARISON

One of the most fundamental tasks in the analysis of molecular high-throughput data sets is the identification of features that show significant differences between two groups. For example, genes with a much higher expression in samples of a diseased group compared to corresponding controls. While group comparison alone is a valuable tool in computational biology, it additionally forms the basis of many downstream analyses, such as the identification of deregulated processes (cf. Section 3.6 + 3.7) or key regulators (cf. Section 3.8). In the following, we introduce several statistical measures and hypothesis tests for this purpose.

3.4.1 *General notation*

All tests in the following paragraph try to determine if two random samples $X = \{x_1, ..., x_n\} \in \mathbb{R}^n$ and $Y = \{y_1, ..., y_m\} \in \mathbb{R}^m$ belong to the same or different populations. To this end, they compare different statistical properties of these samples. Here, we introduce several commonly used sample statistics. The definitions and notations are based on the books "Biostatistical analysis" by Jerrold Zar [603], "100 Statistical Tests" by Gopal K. Kanji [252], "Probability and Statistics" by Morris DeGroot and Mark Schervish [112], and additional sources that are referenced accordingly.

3.4.1.1 *Sample mean*

The arithmetic mean or sample mean of a random sample $X = \{x_1, ..., x_n\}$ is an unbiased estimator of the population mean μ [604]. It is denoted as either $\widehat{\mu}_X$ or $\bar{x}$ and can be calculated as follows:

$$\widehat{\mu}_X = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{20}$$

3.4.1.2 *Sample median*

The sample median of a random sample $X = \{x_1, ..., x_n\}$ is a measure that describes the central value in the sample and that divides all values into an upper half and a lower half. Assuming the sample is sorted, it is defined as:

$$\mathrm{median}(X) = \tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{if } n \text{ is even} \end{cases} \tag{21}$$

3.4.1.3 *Sample variance and standard deviation*

The unbiased variance of a random sample $X = \{x_1, ..., x_n\}$ measures how much the measurements deviate from the sample mean $\widehat{\mu}_X$. It is an unbiased estimator of the population variance σ^2 and can be calculated as follows:

$$\widehat{\sigma}_X^2 = \widehat{Var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu}_X)^2 \tag{22}$$

Since the variance measures deviation from the sample mean with a squared contribution, it is often replaced by the standard deviation, which can be calculated as:

$$\widehat{\sigma}_X = s_X = \sqrt{\widehat{\sigma}_X^2} \tag{23}$$

3.4.1.4 *Sample covariance*

The covariance of two random samples $X = \{x_1, ..., x_n\}$ and $Y = \{y_1, ..., y_n\}$ is a statistical measure describing their linear relationship. The unbiased estimate of the sample covariance is defined as:

$$\widehat{Cov}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu}_X)(y_i - \widehat{\mu}_Y) \tag{24}$$

3.4.2 *Fold-changes*

One of the most popular measures to determine the differences between two measurements x and y in high-throughput experiments is the fold-change. It is generally calculated as the ratio between the two values:

$$fold - change = \frac{x}{y} \tag{25}$$

It can be interpreted as the increase ($fold - change > 1$) or decrease ($fold - change < 1$) of x when compared to y. Since values for increase and decrease are on a different value range, they are hard to compare. For this reason, in practice, a log-ratio is often used instead of the standard ratio:

$$log - fold - change = log(\frac{x}{y}) \tag{26}$$

$$= log(x) - log(y) \tag{27}$$

This transformation has the advantage that the value ranges for increase and decrease are symmetrical and centered around 0, which facilitates their comparison.

3.4.3 *Parametric tests*

Parametric tests are a class of hypothesis tests that assume the data points have a certain underlying distribution.

3.4.3.1 *Welch t-test*

The Welch t-test is a parametric hypothesis test to analyze differences in the mean of two populations with gaussian distribution and unequal variances [571, 572]. Given two independent samples $X = \{x_1, ..., x_n\}$ and $Y = \{y_1, ..., y_m\}$ with sample means $\widehat{\mu}_X$ and $\widehat{\mu}_Y$ and sample standard deviations $\widehat{\sigma}_X$ and $\widehat{\sigma}_Y$ respectively. Then H_0 and H_1 for a two-sided test are defined as follows:

$$H_0 : \mu_X = \mu_Y \tag{28}$$

$$H_1 : \mu_X \neq \mu_Y \tag{29}$$

The test statistic t_ν for the Welch t-test is defined as:

$$t_\nu = \frac{\widehat{\mu}_X - \widehat{\mu}_Y}{\sqrt{\frac{\widehat{\sigma}_X^2}{n} + \frac{\widehat{\sigma}_Y^2}{m}}} \tag{30}$$

A p-value for t_ν can be derived from a t-distribution with ν degrees of freedom. Here, ν can be estimated using the Welch- Saitterthwaite equation [467, 468, 572]:

$$\widehat{\nu} = \frac{(\frac{\widehat{\sigma}_X}{n} + \frac{\widehat{\sigma}_Y}{m})}{(\frac{\widehat{\sigma}_X}{n})^2/(n-1) + (\frac{\widehat{\sigma}_Y}{m})^2/(m-1)} \tag{31}$$

3.4.3.2 *Shrinkage t-test*

The Shrinkage t-test proposed by Opgen-Rhein and Strimmer [398] is a regularized version of the Welch t-test that can be applied if multiple tests are performed simultaneously on a data set, e.g., one test per gene in a gene expression matrix. The main idea of this method is that the individual variance estimates are regularized by shrinking them towards the median. This reduces the effect of outliers and makes the tests more stable. For this reason, the Shrinkage t-test is also suited for data sets with small sample sizes.

Given a data set with p variables and the respective unbiased estimates of the standard deviation $S = \{\widehat{\sigma}_1, ..., \widehat{\sigma}_p\}$. Then the shrinkage estimator can be defined as:

$$\widehat{\sigma}_k^* = (\hat{\lambda}^*)median(S) + (1 - \hat{\lambda}^*)\widehat{\sigma}_k \tag{32}$$

Here $\hat{\lambda}^*$ is the optimal pooling parameter:

$$\hat{\lambda}^* = \min\{1, \frac{\sum_{k=1}^{p} \widehat{Var}(\hat{\sigma}_k)}{\sum_{k=1}^{p} (\hat{\sigma}_k - \text{median}(S))^2}\} \tag{33}$$

$\widehat{Var}(\hat{\sigma}_k)$ is an estimator for the variance of the sample standard deviations. It can be computed using the unbiased variance estimator [398]. Let x_{ik} be the measured value for sample i and variable k and $\bar{x}_{.k}$ the mean value of variable k across all samples. For each x_{ik} the deviation from the mean can then be calculated as:

$$w_{ik} = (x_{ik} - \bar{x}_{.k})^2 \tag{34}$$

Accordingly, let $\bar{w}_{.k}$ be the average deviation from $\bar{x}_{.k}$

$$\bar{w}_{.k} = \frac{1}{n} \sum_{i=1}^{n} w_{ik} \tag{35}$$

Since, $\hat{\sigma}_k = \frac{n}{n-1} \bar{w}_{.k}$, we have:

$$\widehat{Var}(\hat{\sigma}_k) = \frac{1}{(n-1)^3} \sum_{i=1}^{n} (w_{ik} - \bar{w}_{.k})^2 \tag{36}$$

3.4.3.3 *Tests for discrete data*

While t-tests are commonly used to analyze microarray data sets, they are not well suited for RNA-Seq experiments, where raw measurements are discrete counts. Hence a variety of methods have been proposed that are based on discrete probability distributions, like the Poisson distribution (e.g., PoissonSeq[301]) or the negative binomial distribution (e.g., DESeq [20], edgeR [443], and RUVSeq [441]). More information on these methods can be found in the respective publications.

3.4.4 *Non-parametric tests*

In contrast to parametric tests, non-parametric tests make no assumption about the distribution of the analyzed data. Hence, they are more flexible and can be applied in cases where the assumptions of parametric tests, such as t-tests, are violated. However, if the assumptions of parametric tests are met, they can have a higher statistical power than non-parametric ones [98].

3.4.4.1 *Wilcoxon rank-sum test*

The Wilcoxon rank-sum test [440] (WRS test) is a non-parametric hypothesis test that evaluates if two independent random samples are drawn from populations with the same underlying distribution

function. Given two independent samples $X = \{x_1, ..., x_n\}$ and $Y = \{y_1, ..., y_m\}$. Let F_X and F_Y be the distribution functions from which X and Y are drawn. Accordingly, the H_0 and H_1 for a two-sided test can be formulated as follows:

$$H_0 : F_X = F_Y \tag{37}$$

$$H_1 : F_X \neq F_Y \tag{38}$$

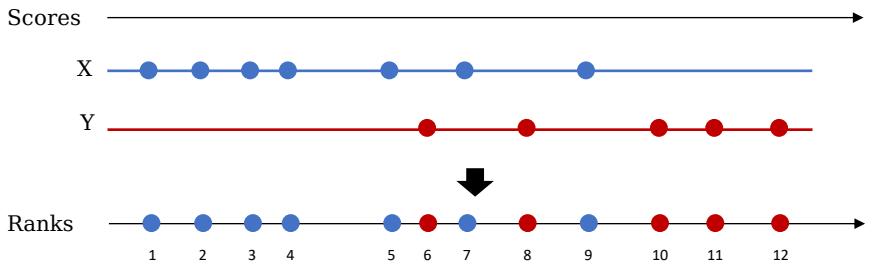


Figure 34: Illustrative example of a Wilcoxon rank-sum test. Shown are the data points of sample X in blue and Y in red. The position of each data point indicates the associated score. In the WRS test the data points are combined and ranked. These ranks are then used to calculate the respective test statistic.

In order to conduct a WRS test, the data points for x and y are combined and sorted increasingly (cf. Figure 34). The test statistic then sums up all ranks of sample x:

$$W_{n,m} = \sum_{i=1}^{n} R(x_i), \tag{39}$$

where R is a function that assigns the rank to each data point.

For large sample sizes ($n > 10$ and $m > 10$) $W_{n,m}$ is approximately normal distributed. Hence, the test statistic is often standardized:

$$Z = \frac{W_{n,m} - \hat{\mu}_W}{\hat{\sigma}_W} \sim N(0, 1) \tag{40}$$

Here, $\hat{\mu}_W$ and $\hat{\sigma}_W$ are estimators for the mean and standard deviation, which are derived from the sample sizes n and m:

$$\hat{\mu}_W = \frac{n(n+m+1)}{2} \tag{41}$$

$$\hat{\sigma}_W = \sqrt{\frac{n \cdot m(n+m+1)}{12}} \tag{42}$$

While for larger sample sizes p-values can directly derived from a standard normal distribution, they have to be looked up in a precomputed table for smaller ones.

3.5 CORRELATION AND DISTANCE MEASURES

A common task in the analysis of molecular high-throughput profiles is to test how similar or dissimilar (or distant) two variables are. In this section, we introduce different statistical measures for this purpose that are used throughout this thesis.

3.5.1 *Distance measures*

Popular approaches to describe if two samples are similar are distance metrics. Given two samples $X = \{x_1, ..., x_n\}$ and $Y = \{y_1, ..., y_n\}$, these metrics interpret the samples as points in an n-dimensional Euclidean space and then use different measures to describe their distance:

The Euclidean distance is defined as the length of a straight line between the two points X and Y:

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{43}$$

For convenience purposes, the square root is in practice sometimes omitted and the squared distance is used instead [505]:

$$d^2(X, Y) = \sum_{i=1}^{n} (x_i - y_i)^2 \tag{44}$$

In contrast to this, the Manhattan distance describes the difference between two points X and Y as the sum of absolute differences in each coordinate:

$$d(X, Y) = \sum_{i=1}^{n} |x_i - y_i| \tag{45}$$

3.5.2 *Correlation coefficients*

Correlation coefficients are statistical measures that describe the association or relationship between two random variables. The absolute value of a correlation coefficient $\rho \in [0, 1]$ indicates the strength of association, where $\rho = 1$ can be interpreted as perfect correspondence and $\rho = 0$ as no relationship. The sign of a ρ indicates the direction of the association: same ($\rho > 0$) or inverse ($\rho < 0$). The different types of correlation coefficients are distinguished by the type of functional relationship they are able to capture.

3.5.2.1 *Pearson's correlation coefficient*

Pearson's correlation coefficient (PCC or Pearson's r) [167, 408] is a statistical measure to quantify the linear dependency between two random samples $X = \{x_1, ..., x_n\}$ and $Y = \{y_1, ..., y_n\}$. It is defined as a normalized version of the sample covariance (cf. Section 3.4.1.4):

$$\widehat{\rho}_{X,Y} = \widehat{\mathrm{Cor}}(X, Y) = \frac{\widehat{\mathrm{Cov}}(X, Y)}{\widehat{\sigma}_X \widehat{\sigma}_Y} \tag{46}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \widehat{\mu}_X}{\widehat{\sigma}_X}\right)\left(\frac{y_i - \widehat{\mu}_Y}{\widehat{\sigma}_Y}\right) \tag{47}$$

3.5.2.2 *Spearman rank correlation coefficient*

The Spearman rank correlation coefficient (r_s) [504] for two samples $X = \{x_1, ..., x_n\}$ and $Y = \{y_1, ..., y_n\}$ is defined as the Pearson correlation between the ranks of the data points instead of their score. Due to the rank transformation step, r_s is able to describe monotone relationships rather then strictly linear ones. Formally, it can be calculated as follows:

$$\widehat{\rho}_{R(X),R(Y)} = \widehat{\mathrm{Cor}}(R(X), R(Y)) = \frac{\widehat{\mathrm{Cov}}(R(X), R(Y))}{\widehat{\sigma}_{R(X)} \widehat{\sigma}_{R(Y)}} \tag{48}$$

where $R(X)$ and $R(Y)$ define the ranks of X and Y.

In case the ranks of all elements in the sample are different, i.e., there are no tied values, it can also be defined as:

$$\widehat{\rho}_{R(X),R(Y)} = 1 - \frac{6 \sum_{i=1}^{n} (R(x_i) - R(y_i))^2}{n(n^2 - 1)} \tag{49}$$

3.5.2.3 *Distance correlation*

The distance correlation is a statistical measure that describes linear and non-linear relationships between two paired vectors $X = \{X_1, ..., X_n\}$ and $Y = \{Y_1, ..., Y_n\}$ with arbitrary and potentially unequal dimensions [133, 519, 520]. To this end, the calculation of the distance correlation is based on pairwise distances between the data points in each vector and not the points directly.
Hence, in a first step, both vectors are individually transformed into n by n matrices that contain all pairwise distances. Accordingly, let a_{ij} and b_{ij} be the distances between data points i and j in vector X and Y respectively.

$$a_{ij} = \|X_j - X_i\| \tag{50}$$
$$b_{ij} = \|Y_j - Y_i\| \tag{51}$$

Here, $\|.\|$ denotes the Euclidean norm.

Next, the pairwise distances are double centered:

$$A_{ij} = a_{ij} - \overline{a_{i.}} - \overline{a_{.j}} + \overline{a_{..}} \tag{52}$$
$$B_{ij} = b_{ij} - \overline{b_{i.}} - \overline{b_{.j}} + \overline{b_{..}} \tag{53}$$

Here, $\overline{a_{i.}}, \overline{b_{i.}}$ denote the means of row i, $\overline{a_{.j}}, \overline{b_{.j}}$ the means of column j, and $\overline{a_{..}}, \overline{b_{..}}$ the mean across all pairwise distances in sample x and y.

Using the double centered distances, we can now define the following estimator for the squared distance covariance [519, 520]:

$$\widehat{dCov}^2(X, Y) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} B_{ij} \tag{54}$$

Analogously, the distance variance can be calculated as the distance covariance of two identical vectors:

$$\widehat{dVar}(X) = \widehat{dCov}(X, X) \tag{55}$$

Finally, the distance correlation is defined as the distance covariance scaled by the individual distance variances [133]:

$$\widehat{dCor}(X, Y) = \frac{\widehat{dCov}(X, Y)}{\sqrt{\widehat{dVar}(X)\widehat{dVar}(Y)}} \tag{56}$$

3.6 ENRICHMENT ANALYSIS

After normalization and quality control, one of the next steps in the analysis of molecular high-throughput data is generally some sort of feature selection or feature prioritization. To this end, researchers usually identify biological features that either differentiate between two groups (cf. Section 3.4), or that correlate with a phenotype (cf. Section 3.5), such as the disease status or treatment response. Hereafter, we call these features test set.

A very popular next step is to investigate if the features in our test set are involved in deregulated biological processes or signaling pathways. For this purpose, various methods have been proposed that test if predefined sets of features, i.e., biological categories, are significantly enriched in a test set. We refer to this process as "enrichment analysis", although other terms are also regularly used in literature, such as "gene set analysis", "gene set enrichment analysis", or "pathway analysis".

In the following paragraphs, we first describe the general structure of an enrichment analysis workflow as described by Ackermann and Strimmer [2]. Subsequently, we give an overview of different enrichment analysis approaches that are relevant for this thesis.

3.6.1 *General structure of an enrichment analysis workflow*

Due to their popularity, enrichment analysis methods have been extensively studied and many studies review, compare and discuss properties, performance, and challenges of the different approaches (see e.g., [2, 226, 231, 264, 374, 376, 501]). Here, we give an overview of the work by Ackermann and Strimmer [2]. In their paper, the authors compare and discuss the theoretical properties of different approaches and develop a "general modular framework for enrichment analysis" [2]. Within this framework, they describe two distinct strategies: modular workflows and global tests. Both start with a matrix of measured features in different samples. In global approaches, a multivariate statistic is applied to the entire input matrix at once, while in the modular strategy different processing steps are performed to identify enriched biological processes. In this section, we focus on the latter, since this is the main use case of our GeneTrail web service (cf. Chapter 5).

In general, the modular strategy is divided into four distinct steps that are described in the following: (1) feature-level statistic, (2) score transformation, (3) set-level statistic, and (4) significance assessment. An overview of this workflow is depicted in Figure 35.

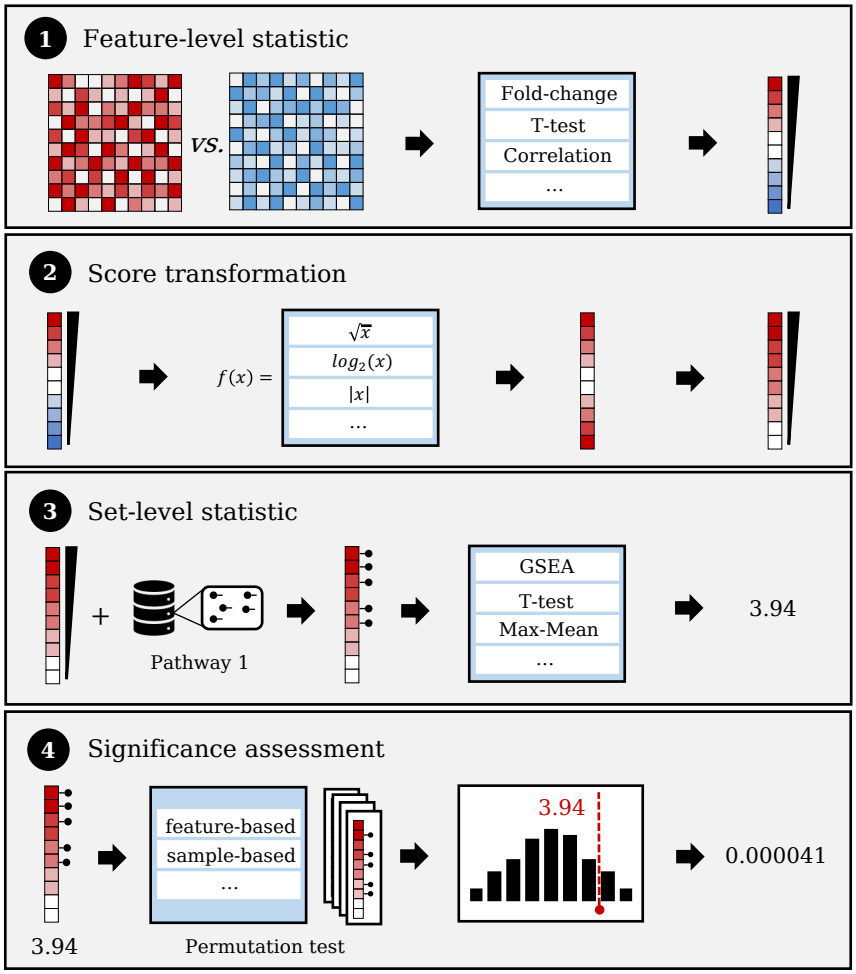


Figure 35: General modular structure of an enrichment analysis workflow as described by Ackermann and Strimmer [2]. (1) Feature-level statistic, (2) score transformation (optional), (3) set-level statistic, and (4) significance assessment via permutation test.

3.6.1.1 *Feature-level statistic and score transformation*

The goal of the feature-level statistic (or "gene-level statistic" [2]) is to identify molecular features that show significant differences between two sample groups, e.g., disease and control. To this end, any method for group comparison can be applied, such as the statistical tests described in Section 3.4.

In the next step, the resulting scores can optionally be transformed, e.g., by using logarithmized values to reduce the effect of outliers, or by using absolute values to ignore the sign of the feature-level statistic.

3.6.1.2 *Set-level statistic and significance assessment*

The score list is then used to analyze the enrichment of a predefined biological category, such as biological processes from GO [100] or signaling pathways from KEGG [394] (cf. Section 3.2.3). For this purpose, a "set-level statistic" is applied that for each biological category tests if it is significantly enriched or depleted in the analyzed score list. To this end, various statistical tests can be used. An overview is provided in the Sections 3.6.2-3.6.3.3.

In order to assess if the set-level statistic is statistically significant a p-value is calculated. To this end, Ackermann and Strimmer describe three different strategies: a feature-based strategy, a sample-based one, and a combination of both called restandardization [136].
In the sample-based strategy, a p-value for a test statistic t is defined as the proportion of randomly permuted sample labels that lead to equal or more extreme values.
For the restandardization approach, both types of permutations are combined and the resulting test values are normalized with respect to the mean and standard deviations across all conducted permutations runs before a p-value is estimated [136].
For the feature-level strategy, the p-values can either be calculated exactly (see e.g., [259]), by an approximation of the probability distribution (cf. Section 3.4), or by using a permutation test (see subsequent paragraph). For both, the sample-based strategy and restandardization approach, a p-value can only be calculated using a permutation test.

Permutation test

Given a test statistic t and a number of random permutations $\hat{t}_1, ..., \hat{t}_n$ generated using either strategy, an empirical upper-tailed p-value can be calculated as:

$$P(T \geqslant t) = \frac{\sum_{i=1}^{n} I(\hat{t}_i \geqslant t)}{n}, \tag{57}$$

where I is an indicator function that determines if a logic expression b is true or false and returns the values 1 and 0 respectively:

$$I(b) = \begin{cases} 1 \text{ if } b \\ 0 \text{ else} \end{cases} \tag{58}$$

In practice, a pseudo-count is often added to the above formula to avoid zero values, which might cause problems for a small number of permutations [271].

$$P(T \geqslant t) = \frac{1 + \sum_{i=1}^{n} I(\hat{t}_i \geqslant t)}{n} \tag{59}$$

Empirical lower-tailed and two-sided p-values can be calculated by modifying the inequality accordingly.

3.6.2 *Over-representation analysis (ORA)*

One of the most general enrichment analysis methods is the so-called over-representation analysis (ORA) [127]. This method checks for any given biological category if a test set has more entries in this category than expected based on a reference (or background) set. In this context, the background set usually includes all features that are measured, or could potentially be measured in a high-throughput experiment, while the test set often only comprises a small subset, such as the most up-regulated genes.

An ORA is often modelled as a standard urn experiment that tests if a certain biological category is over-represented (enriched) or under-represented (depleted) in the test set. An example for this model is shown in Figure 36.

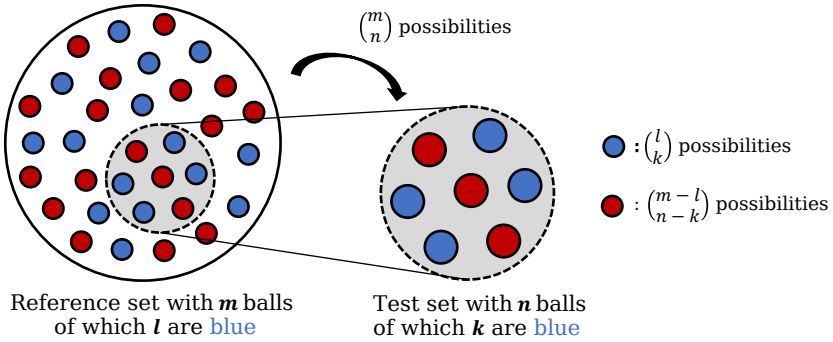


Figure 36: Model used for the over-representation analysis (ORA). We have an urn (reference set) with m balls of which l are blue and the rest are red. From the urn we draw n balls of which k are blue.

Here, we describe the version of ORA that was proposed by Backes et al. for the original GeneTrail web service [29]. Assume we are given a test set $T = \{t_1, ..., t_n\}$, a reference set $R = \{r_1, ..., r_m\}$ and a biological category $C = \{c_1, ..., c_l\} \subset R$ with k entries in T. Then, we would expect to find $k' = \frac{n \cdot l}{m}$ entries of category C in our test set. To judge, if the difference between k and k' is significant, we can calculate an one-sided p-value for C. If $k > k'$ a upper-tailed p-value is calculated

and otherwise a lower-tailed one.

If $T \subset R$, the p-value k matches is calculated using the hypergeometric test:

$$P_C = \begin{cases} \sum_{i=k}^{n} \frac{\binom{l}{i}\binom{m-l}{n-i}}{\binom{m}{n}} & \text{if } k' < k \\ \sum_{i=0}^{k} \frac{\binom{l}{i}\binom{m-l}{n-i}}{\binom{m}{n}} & \text{if } k' \geqslant k \end{cases} \tag{60}$$

If $T \not\subset R$, the Fisher's exact test is used instead:

$$P_C = \begin{cases} \sum_{i=k}^{n} \frac{\binom{n}{i}\binom{m}{l+k-i}}{\binom{m+n}{l+k}} & \text{if } k' < k \\ \sum_{i=0}^{k} \frac{\binom{n}{i}\binom{m}{l+k-i}}{\binom{m+n}{l+k}} & \text{if } k' \geqslant k \end{cases} \tag{61}$$

3.6.3 *Functional class scoring (FCS)*

In contrast to ORA-based approaches that consider only a subset of the measured features, e.g., the features with the highest degree of deregulation, Functional Class Scoring (FCS) methods [264] use all measured features to analyze if a biological category C is significantly enriched or depleted.

3.6.3.1 *Gene set enrichment analysis (GSEA)*

One of the most popular functional class scoring method is the gene set enrichment analysis (GSEA) [29, 514]. This method is based on the Kolmogorov-Smirnov test, which analyzes if two populations have the same distribution [274, 496]. In the context of enrichment analysis, this test can be employed to determine if a biological category is significantly enriched in the beginning or the end of a sorted feature list.

Over the years, two different versions of GSEA have been established: a weighted version [514] and an unweighted one [29, 259]. Both are described in the following paragraphs.

Weighted GSEA

The weighted version of GSEA tests for a decreasingly sorted list $L = [l_1, ..., l_n]$, if the entries of a biological category $C = \{c_1, ..., c_j\} \subset L$ are significantly enriched in the beginning or end of L, i.e., if they are enriched or depleted. To this end, a running-sum statistic RS is calculated by iterating through the sorted list, from the largest element to

the smallest, and if the list element at position i belongs to C, RS is increased, otherwise RS is reduced:

$$RS[i] = \begin{cases} 0 & \text{if } i = 0 \\ RS[i-1] + \frac{|w(l_i)|^P}{N_R} & \text{if } l_i \text{ in } C \\ RS[i-1] - \frac{1}{(n-j)} & \text{else} \end{cases} \tag{62}$$

Here, $w(l_i)$ is the score of the i^{th} element in l, $N_R = \sum_{l_i \in C} |w(l_i)|^P$ the sum of all elements in l that belong to C, and P is a parameter that controls the influence of the used scores. By using this weighting scheme, the running-sum statistic starts and ends at 0.

The test statistic (enrichment score ES_C) is then defined as the element in RS with the maximum deviation from 0:

$$ES_C = \max_{i \in [1,n]} \{|RS[i]|\} \tag{63}$$

Finally, a p-value for ES_C is calculated using a permutation test.

Unweighted GSEA

In contrast to the weighted version, the unweighted GSEA solely relies on the order of a sorted list $L = [l_1, ..., l_n]$ in order to determine the score for a category $C = \{c_1, ..., c_j\} \subset L$. A huge advantage of the unweighted version is that we do not need to perform a permutation test to calculate a p-value. Instead, we can use a dynamic programming approach to calculate an exact p-value [259].
Analogously to the weighted GSEA, a running-sum statistic is employed to calculate the final test statistic ES_C:

$$RS[i] = \begin{cases} 0 & \text{if } i = 0 \\ RS[i-1] + (n-j) & \text{if } l_i \text{ in } C \\ RS[i-1] - j & \text{else} \end{cases} \tag{64}$$

$$ES_C = \max_{i \in [1,n]} \{|RS[i]|\} \tag{65}$$

In this approach, the value of $RS[i]$ is j times increased by $n-j$ and $n-j$ times decreased by j. Hence, similar to the weighted version, the running-sum statistic $RS[i]$ starts and ends at 0. Since the values in which $RS[i]$ is modified in each step are constant, the final score of ES_C is only determined based on the sequence of increases and decreases, i.e., a particular permutation of list L. It has even been observed that all values of $RS[i]$ can be calculated using a grid of size $(n-j) \times j$, where the points in the grid determine all possible values of $RS[i]$ [259] (cf. Figure 37). Accordingly, each path in this grid constitutes one possibility to calculate a particular enrichment score.

Consequently, the p-value for ES_C can be defined as the proportion of paths in the grid, i.e. the proportion of permutations of L, that have a score of at least ES_C [259].

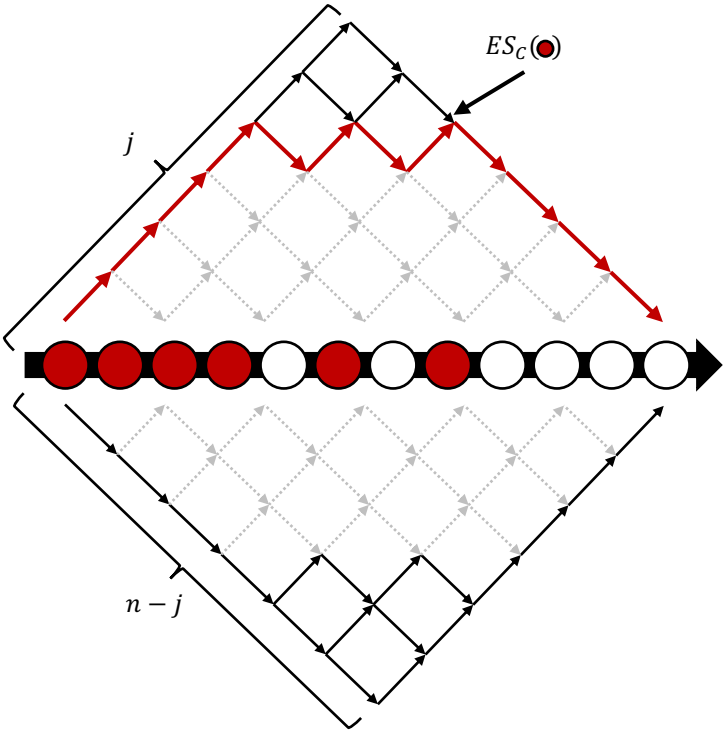


Figure 37: Example of a running-sum statistic. Red balls indicate genes that belong to a biological category. The points in the grid indicate all possible values that can be achieved by the running sum statistic, and the arrows mark all possible paths in this grid. Red arrows mark the path of the running sum for this example. Arrows with a solid line represent all possible paths with the same or a higher deviation from zero.

Based on these observations, Keller et al. [259] developed an algorithm that calculates the exact p-value for an enrichment score ES_C via the complement of this event:

$$p = 1 - \frac{X}{Y} \tag{66}$$

Here, X is defined as the number of paths in the grid with a score smaller than ES_C, and $Y = \binom{n}{j}$ as the total number of paths. Consequently, $1 - \frac{X}{Y}$ denotes the probability to obtain a score of at least ES_C.

For the calculation of X, Keller et al. [259] propose the following dynamic programming algorithm over a matrix $M^{(j+1)x(n-j+1)}$. Each entry $M[i,k]$ of M indicates the number of paths with i members of a biological category and $k-i$ non-members that have an absolute enrichment score smaller than $|ES_C|$.

The first row and column of M are initialized with:

$$M[i,0] = \begin{cases} 1 & \text{if } -|ES_C| < i \cdot (n-j) < |ES_C| \\ 0 & \text{else} \end{cases} \tag{67}$$

$$M[0,k] = \begin{cases} 1 & \text{if } -|ES_C| < -k \cdot j < |ES_C| \\ 0 & \text{else} \end{cases} \tag{68}$$

For $i \in [1,j]$ and $k \in [1,(n-j)]$ the remaining entries $M[i,k]$ can then be computed using the following recurrence:

$$M[i,k] = \begin{cases} M[i-1,k] + M[i,k-1] & \text{if } -|ES_C| < (*) < |ES_C| \\ 0 & \text{else} \end{cases}, \tag{69}$$

where $(*)$ is defined as $i \cdot (n-j) - k \cdot j$.

The dynamic programming approach described above is used to calculate two-sided p-values. For the calculation of one-sided p-values, the inequalities need to be restricted to one side.

3.6.3.2 *Further hypothesis tests for group comparison*

In addition to the Kolmogorov-Smirnov test (GSEA), many other hypothesis tests for group comparison can also be applied to detect deregulated biological categories, such as the Wilcoxon rank-sum test and the Welch t-test (cf. Section 3.4).
Given a sorted list $L = [l_1, ..., l_n]$ with n entries of which j belong to a biological category $C = \{c_1, ..., c_j\} \subset L$, these methods are applied to compare the category members in the list against the non-category members.

3.6.3.3 *Averaging methods*

Next to hypothesis tests, a further class of enrichment analysis approaches are so called averaging methods [2]. These approaches calculate a sample statistic, such as the mean or median (cf. Section 3.4), for all features in the test set that belong to a specific category. The significance of these scores is assessed via a permutation test.

However, this can be problematic if a category contains both positive an negative scores that might cancel each other out. To address this issue, Efron and Tibshirani propose a new sample statistic called max-mean, which calculates the mean of positive and negative values separately and then uses the absolute maximum of both as a test statistic [136].

Ackermann and Strimmer have shown that averaging methods have a similar and in some instances better performance than the methods described in the previous sections [2].

3.7 NETWORK ANALYSIS

The enrichment analysis methods described in the last section utilize predefined sets of biological features (genes, proteins, or miRNAs) to identify deregulated biological processes. However, some of the databases, from which these categories are extracted, provide additional information about the relationship between the different features. In order to make use of this information, network-based (or topology-based) methods have been developed that utilize the graph structure provided by databases like KEGG [394] to infer deregulated biological processes.
To this end, the graph topology is usually combined with molecular measurements from high-throughput experiments that are processed to obtain weights for vertices or edges of the analyzed network. The general structure of a network analysis workflow is depicted in Figure 38.

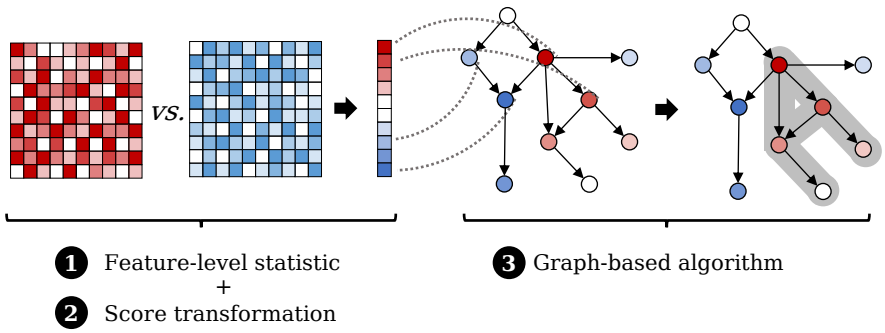


Figure 38: General structure of a network analysis workflow. Scores obtained from a group comparison (1+2) are mapped onto the network structure. (3) A graph-based algorithm is applied to identify deregulated subgraphs.

3.7.1 *Subgraph ILP*

For the identification of deregulated pathways in a biological network, Backes et al. propose an integer linear-programming (ILP) algorithm [32]. Next to the graph topology, this approach also requires scores for the molecular features in the graph, e.g., gene expression differences obtained from a group comparison of diseased and healthy samples. These scores are then used as vertex weights in the network. Given the weighted network, the method searches for the heaviest

connected subnetwork that can be reached from a designated root node. Here, the root is assumed to be a key player in the identified network, e.g., a potential cause or regulator for the observed alterations. A formal definition of the problem is described in the following section.

ILP formulation

Given a directed and weighted graph $G = (V, E)$ with n vertices $\{v_1, ..., v_n\} \in V$, the ILP searches for the connected subgraph $S \subset V$ of size k that has the highest (absolute) weight and can be reached from one designated root node.
Let $x_i \in \mathbb{B}$ and $y_i \in \mathbb{B}$ be two binary decision variables for each node v_i, where x_i specifies if a vertex is selected for the final solution and y_i indicates if v_i is the root of the selected subgraph. Additionally, let w_i be the weight of a vertex v_i. Then the objective function of the ILP can be defined as:

$$\max_{x \in \mathbb{B}} \sum_{i=1}^{n} w_i x_i \tag{70}$$

Additionally, various constraints are required to find a correct solution. The first constraint makes sure that only subgraphs of size k are selected.

$$\sum_{i=1}^{n} x_i = k \tag{71}$$

The next two inequalities guarantee that only one root is selected and that it is part of the selected solution.

$$\sum_{i=1}^{n} y_i = 1 \tag{72}$$

$$y_i \leqslant x_i \quad \forall i \in [1, n] \tag{73}$$

Moreover, we need to ensure that the selected subgraph is connected, i.e., that a vertex is either the root or the target of a vertex in our solution. Let $In(i)$ define the indices of all nodes that target v_i. Then, we can define the new constraint as:

$$x_i - y_i - \sum_{j \in In(i)} x_j \leqslant 0 \quad \forall i \in [1, n] \tag{74}$$

Since this constraint is also satisfied by multiple disconnected cycles, an additional inequality is necessary to exclude this case. Let C be the node indices of a particular cycle and $In(i)$ the indices of all vertices with outgoing edges that target any vertex in cycle C, then the additional constraint can be formulated as follows:

$$\sum_{i \in C} (x_i - y_i) - \sum_{j \in In(C)} x_j \leqslant |C| - 1 \quad \forall C \tag{75}$$

3.8 REGULATOR IMPACT ANALYSIS

In the previous sections, we presented different methods for the analysis of differentially expressed genes as well as deregulated biological processes. A logical next step is the identification of the key regulatory factors that might control these processes.

In this context, especially algorithms for the analysis of transcriptional regulators like transcription factors, co-factors and chromatin modifiers have been discussed. These methods can be categorized into two classes: RTI-based approaches and Motif-based ones. Examples for both are presented in the following.

3.8.1 *RTI-based approaches*

The first class of approaches uses experimentally determined regulator-target gene interactions (RTIs, cf. 3.2.4.1) to score regulators either solely based on the used collection of RTIs or additionally using different molecular measurements.

3.8.1.1 *Correlation set analysis (CSA)*

Huang et al. presented a method, called correlation set analysis (CSA) [225], that determines the effect of regulators based on the co-expression of their target genes. To this end, for each regulator, the absolute pairwise correlations between its target genes are investigated.

The inputs for CSA are (1) a set T of genes we are interested in, e.g, the most deregulated genes in a group comparison, (2) a matrix X containing gene expression measurements of multiple samples for all genes in T, and (3) a set R of regulators with at least two targets in T.

Given a regulator $R_i \in R$ with n target genes in T and the corresponding gene expression measurements $X = \{X_1, ..., X_n\}$ the following test statistic can be defined:

$$\mathrm{CSA}(R_i, X) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |\widehat{\rho}_{X_i, X_j}| \tag{76}$$

Here, the authors make the assumption that all target genes are similarly affected by a regulator and that the expression measurements of each target pair should either be correlated or anti-correlated. Hence, a large value of $\mathrm{CSA}(R_i, X)$ should indicate a large influence of R_i on the expression of the analyzed genes. The significance of each result is assessed via a permutation test that randomly permutes the assigned target genes of each regulator.

3.8.1.2 *Regulatory impact factors (RIF1 + RIF2)*

Further approaches are the regulatory impact factors (RIF1 and RIF2) proposed by Reverter et al. [437] that determine transcriptional regulators with different effects in two biological conditions, e.g., regulators that are affected by a mutation in one of the groups [230]. To this end, both RIF1 and RIF2 analyze the influence of regulators using (1) the expression values of regulators and target genes in two groups, and (2) the correlation of regulators and their target genes.

Given gene expression measurements for two groups G1 and G2, where a_j is the average expression of a gene j across both groups (abundance), $e1_j$ and $e2_j$ constitute the average expression of gene j in group G1 or G2 respectively, d_j is the difference in expression between G1 an G2 for gene j, and $\widehat{\rho 1}_{ij}$ and $\widehat{\rho 2}_{ij}$ represent the correlation between a regulator i and its target j in the respective group. Based on this information, Reverter et al. define two measures for scoring regulators with altered co-expression patterns.

RIF1 assigns high scores to regulators with highly abundant and highly deregulated target genes that additionally exhibit large differences in co-expression between the two groups.

$$\mathrm{RIF1}(i) = \frac{1}{n}\sum_{j=1}^{n} a_j \times d_j \times (\widehat{\rho 1}_{ij} - \widehat{\rho 2}_{ij})^2 \tag{77}$$

RIF2 assigns high scores to the regulators with the most altered co-expression in the two groups.

$$\mathrm{RIF2}(i) = \frac{1}{n}\sum_{j=1}^{n} \left[(e1_j \times \widehat{\rho 1}_{ij})^2 - (e2_j \times \widehat{\rho 2}_{ij})^2\right] \tag{78}$$

3.8.1.3 *ORA-based approaches*

Over the years, different approaches have been proposed that investigate if the target genes of a certain regulator are over-represented or enriched in a given test set. Hence, we refer to them as ORA-based approaches.

These approaches can be formulated similar to the over-representation analysis described in Section 3.6.2. Assume we are given a regulator with l targets in a background set $R = \{r_1, ..., r_m\}$ and k targets in a test set $T = \{t_1, ..., t_n\}$, then different measures can be applied:

Essaghir et al. [141] use a method called TFactS to identify influential regulators with more targets then expected using the hypergeometric

distribution:

$$P(K \geqslant k) = \sum_{i=k}^{n} \frac{\binom{l}{i}\binom{m-l}{n-i}}{\binom{m}{n}} \text{ (assuming } T \subset R \text{)} \tag{79}$$

In cases where $T \not\subset R$, the Fisher's exact test (cf. Section 3.6.2) can also be applied.

In a similar approach, Yang et al. [593] propose to use a binomial distribution model:

$$P(K \geqslant k) = \sum_{i=k}^{l} \binom{l}{i} \left(\frac{n}{m}\right)^{i} \left(1 - \frac{n}{m}\right)^{n-i} \tag{80}$$

Additionally, Yang et al. [593] use a clustering coefficient based measure to calculate the ratio of targets in the test set compared to the targets in the reference:

$$TDD(i) = \frac{2k}{n(n-1)} \tag{81}$$

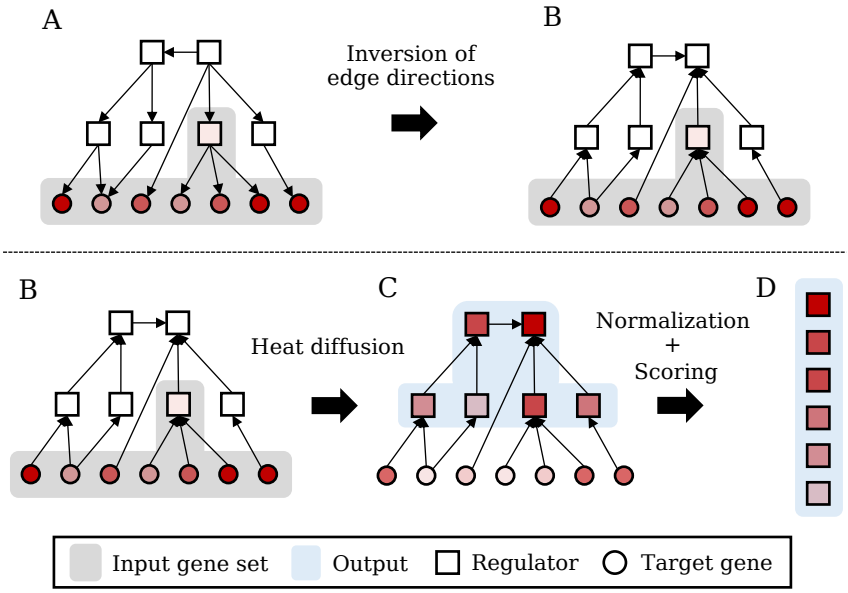


Figure 39: Overview of the TFRank algorithm. (A→B) The network is initialized with scores for each gene in the test set and the edges of the network are inverted. (B→C) Heat diffusion is applied to the inverted RTI network to propagate the scores from target genes (circles) to regulators (squares). (C→D) The regulators are then normalized and ranked.

3.8.1.4 *TFRank*

Regulator-target interactions (RTIs) can also be interpreted as a network, where regulators and target genes represent vertices and each RTI represents a directed edge from the regulator to the target gene. Based on this idea, Goncalves et al. [184] present TFRank, a graph-based algorithm to prioritize transcriptional regulators using heat diffusion (cf. Figure 39).

The inputs for TFRank are a collection of RTIs and a set of relevant genes with corresponding scores, e.g., the most upregulated genes in a comparison between diseased and healthy samples.
For this approach, the collection of RTIs is interpreted as a directed graph $G = (V, E)$, where nodes constitute regulators and genes, and edges a regulator-target interactions. However, the direction of each RTI is inverted, such that it points from the target to the regulator. Moreover, the scores of the test set are used as vertex weights in the graph. The goal of TFRank is now to explore all paths in the network that involve the genes of our test set. To this end, the scores are propagated from the genes of interest to their putative regulators in a process called heat diffusion [184], which formally can be defined as follows.
Let $A^{|V|\times|V|}$ be the adjacency matrix representing the inverted graph, I a unit matrix with the same dimensions, and D a corresponding diagonal matrix, where each diagonal entry specifies the number of outgoing edges. Additionally, let $T \subset V$ be the test set, and p_0 a vector that contains the initial scores for each vertex in the network, such that all vertices that are part of T are assigned the respective score and all others are set to 0.
The heat diffusion is defined iteratively, where in each step the vertex scores are partially transmitted to their direct neighbors. In the first step the initial weights p_0 are propagated as follows:

$$p_1 = p_0 \left(I - t(I - D^{-1}A)\right) \tag{82}$$

Here, t is the heat diffusion coefficient that controls how much information is transmitted. The matrix $D^{-1}A$ describes the transition probability for a random walk in G [184].
A general version of this formula can be defined accordingly:

$$p_{n+1} = p_n \left(I - t(I - D^{-1}A)\right) \tag{83}$$

Finally, the regulators are ranked based on their vertex score. To this end, the heat diffusion scores can either be used directly or normalized, for example by scaling the score for each regulator based on its number of target genes.

3.8.2 *Motif-based approaches*

The RTI-based approaches described in the previous paragraphs rely on experimentally determined binding sites of transcriptional regulators. These binding sites are often only available for well-studied tissue types and, hence, might be incomplete. To overcome this issue, different methods have been proposed that use DNA binding motifs of regulators, e.g., extracted from associated ChIP-seq experiments, to estimate potential binding sites (cf. Section 3.2.4.2). In the following sections, we assume that all motifs are provided as position weight matrices (PWM, cf. Section 3.2.4.2) and that a pseudo-count ($\pi = 1$) is added to each matrix element to avoid values of zero [79, 447].

3.8.2.1 *TRAP*

TRAP uses a biophysical model to estimate the binding affinity of a transcription factor to a specific DNA sequence [329, 447].

For this purpose, the authors assume the binding of a regulator R to any sequence S takes place at an equilibrium $R + S \Leftrightarrow R \cdot S$. Under this assumption, the affinity that R binds to S can be defined as the fraction of bound sites:

$$a = \frac{[R \cdot S]}{[S][R \cdot S]} = \frac{R_0 \cdot e^{-\beta E(S)}}{1 + R_0 \cdot e^{-\beta E(S)}}, \tag{84}$$

where R_0 is a motif-specific constant and $\beta E(S)$ is the mismatch energy at site S [46].

Given a motif-matrix M with W columns and an arbitrary sequence with the same length (S_W). Then $M[i, \alpha_i]$ denotes the entry in column i that corresponds nucleotide $\alpha_i \in [A, C, G, T]$ in S_W. Accordingly, $M[i, max]$ indicates the nucleotide with the largest score in column i. Then the mismatch energy βE for sequence S_W can be calculated as:

$$\beta E(S_W) = \frac{1}{\lambda} \sum_{i=1}^{W} \log \left(\frac{M[i, max]}{M[i, \alpha_i]} b_{i,\alpha_i} \right), \tag{85}$$

where λ is a scale parameter and b_{i,α_i} is the relative background frequency of α_i with respect to most frequent base at position i [46, 447].

For a motif M of length W and a sequence S of length $L > W$ the affinity can then be calculated as:

$$\tilde{a}(S, M) = \sum_{i=1}^{L-W} \frac{R_0 \cdot e^{-\beta E(S[i,i+W-1])}}{1 + R_0 \cdot e^{-\beta E(S[i,i+W-1])}} \tag{86}$$

Finally, for double stranded DNA, the total affinity is calculated as the sum of the individual affinities for S and its reverse complement $\bar{S}$:

$$a_S = \tilde{a}(S, M) + \tilde{a}(\bar{S}, M) \tag{87}$$

Based on ChIP-Seq experiments, Roider et al. [447] screened the parameter space for both λ and R_0. They found that $\lambda = 0.7$ and $R_0 = 0.58 \cdot W - 5.66$ are good estimates that are suitable for most practical applications [447].

3.8.2.2 *TEPIC + INVOKE*

Building upon the idea of TRAP, Schmidt et al. [472] created the TEPIC framework, which combines the transcription factor affinities with epigenetic measurements, such as open-chromatin data, to improve the prediction of gene expression.

To this end, the authors extend the definition of affinities with features extracted from open-chromatin regions. First of all, the binding affinity of a regulator for a specific gene g is only calculated based on the sequence of all open-chromatin regions assigned to this g. For this purpose, Schmidt et al. distinguish between two approaches. In the first approach, the TRAP affinities for all open-chromatin region within a $3,000$bp (3kbp) window around the transcription start site (TSS) are summed up:

$$a_g^{3kb} = \sum_{p \in P_{g,3000}} a_p, \tag{88}$$

where $P_{g,3,000}$ is the set of all open chromatin regions within the $3,000$bp window.

For the second approach, a $50,000$ bp window is used, but the affinity of each open-chromatin region is weighted with the distance to the TSS using an exponential decay function:

$$a_g^{50kb} = \sum_{p \in P_{g,50,000}} a_p \cdot e^{-\frac{d_{p,g}}{d_0}}, \tag{89}$$

where $d_{p,g}$ is the mean distance from the middle of the open-chromatin region to the TSS and $d_0 = 5,000$bp a normalization factor.
Additionally, in both approaches the affinity of each open-chromatin region can additionally be weighted with the abundance s_p of each peak $p \in P_g$.

$$a_g^{3kb-S} = \sum_{p \in P_{g,3,000}} s_p \cdot a_p \tag{90}$$

$$a_g^{50kb-S} = \sum_{p \in P_{g,5,0000}} s_p \cdot a_p \cdot e^{-\frac{d_{p,g}}{d_0}} \tag{91}$$

An overview of the whole approach is depicted in Figure 40.

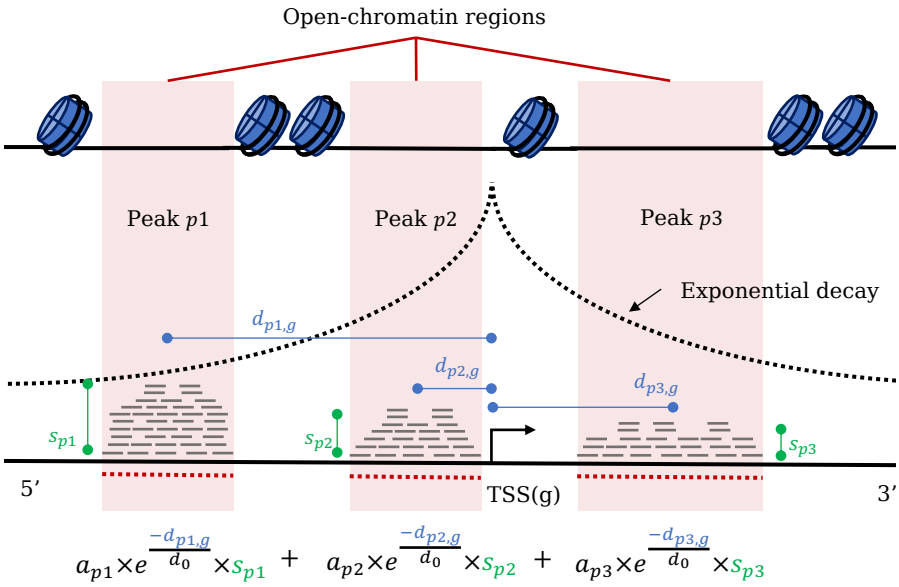


Figure 40: Overview of the TEPIC algorithm. Affinities for a specific regulator are only calculated for open-chromatin regions assigned to a gene (marked in red). The affinities of each open-chromatin region is optionally weighted with the signal abundance of the peak as well as the weighted distance to the TSS (exponential decay).

INVOKE

Schmidt et al. showed that the calculated affinities can also be used to predict gene expression [472]. Hence, in addition to open-chromatin regions of a particular biological sample, this approach also requires corresponding gene expression values as input. Using both data types, the authors use the following approach to find the set of regulators that are good predictors for the provided gene expression values. First, TEPIC is applied to the open-chromatin regions to calculate binding affinity scores for each regulator and target gene, as described in the previous section. Then, given the expression values as our response y and the predicted affinity of all regulators and genes as feature matrix X, the following linear regression model with elastic net penalty can be trained:

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2 + \alpha\|\beta\|_2^2 + (1-\alpha)\|\beta\|_1, \tag{92}$$

where β are the feature coefficients and $\alpha \in [0, 1]$ a weight to control the ratio between ridge and lasso penalty. For a detailed description of the elastic net regularization, see [208].

In the end, each $\hat{\beta}[i]$ indicates the importance of regulator i in the model. Hence, they can directly be used to identify and prioritize the most influential ones. The elastic net penalty in the model helps to control the influence of highly correlated regulators.

4

GRAVITON

The analysis of molecular high-throughput profiles often is a multi-step process that involves different computational tools and a variety of external databases. Hence, all resources for the pipeline have to be carefully selected. Computational approaches must be adapted to the properties of the analyzed data set, external databases must be carefully curated, and identifiers for genes or proteins for all data sets need to be sanitized and mapped into a uniform representation (cf. Section 4.3.2). On top of this, these workflows often are computationally expensive and require efficient algorithms as well as powerful hardware with sufficient resources that might not be accessible to everybody.

For these reasons, bioinformatics pipelines are often implemented as web services. For users, this ensures that all components are compatible, and it removes all technical challenges, like installing dependencies or processing third-party databases. Consequently, they can spend more time on the analysis of their data set.

In this chapter, we first introduce the technologies that build the foundations of modern web services. Subsequently, we introduce the general framework that we developed to implement all web services discussed in this thesis (cf. Figure 41).

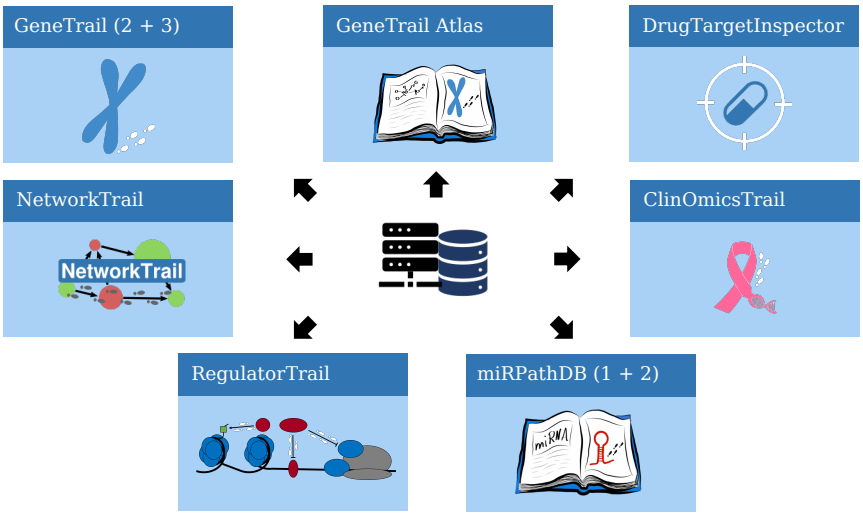


Figure 41: Overview of web services and databases build using the Graviton framework. GeneTrail, RegulatorTrail, and miRPathDB are described in this thesis, see Chapters 5, 8, and 6. DrugTargetInspector [474] and ClinomicsTrail [475] are described in Appendix C.

4.1 THE FUNDAMENTALS OF MODERN WEB APPLICATIONS

The foundations of most modern web services are interactions between client devices and a web server via standardized protocols. In general, a web service manages different resources that the client can address, query, or modify. For this purpose, web services often provide an application programming interface (API) that defines a set of available operations for external users.
In the following paragraphs, we introduce the web technologies used to create our web services. First, we define how information provided by web services on the internet can be addressed (cf. Section 4.1.1). Then, we describe the communication protocols that are supported by our server (cf. 4.1.2). Finally, we introduce the REST architectural style that was used as a guideline to create our web framework (cf. 4.1.3).

4.1.1 *Resources and Uniform Resource Identifier (URI)*

A resource in the context of the World Wide Web (WWW) [49] is any abstract or physical object that can be addressed via the internet, e.g., documents, websites, or even actions performed by a server. To this end, each resource is clearly defined by a unique Uniform Resource Identifier (URI) [48, 49].
URIs are short strings that uniquely encode resources on a network, such as the internet [48]. In general, a URI consists of five distinct parts [50]: (1) The scheme, which defines the method that should be used to address the resource, e.g., the network protocol (cf. Section 4.1.2). (2) The authority that manages the resource, i.e., the host. (3+4) A path of text segments, which are separated by "/", and a query string that both identify the resource. (5) The fragment that references a specific part of a resource, like a paragraph on a website. The generic structure of a URI and a real-world example are depicted in Listing 1.

In case a URI additionally provides information about the location of a resource on the server, it is referred to as a Uniform Resource Locator (URL) [51].

Listing 1: Generic structure of a URI adapted from [580] and real world example. Elements in brackets are optional.

```
scheme:[//authority]path[/?query][/#fragment]
scheme:[//[user@]host:port]path[/?query][/#fragment]

http://dti.bioinf.uni-sb.de/help?topic=p_value_adjustments/#FDR
```

4.1.2 *The Internet Protocol Suite (TCP/IP)*

The communications between different devices on the internet are enabled by a set of highly standardized protocols, called the internet protocol suite (TCP/IP), which specifies how data on the internet is addressed, transferred, and received [72, 73, 87]. Here, we focus on the application layer protocols HTTP and HTTPS that are used for the client-server communications with our web services.

4.1.2.1 *Hypertext Transfer Protocol (HTTP)*

The Hypertext Transfer Protocol (HTTP) [149] is a request-response protocol for the stateless data transfer on a network, such as the internet. In general, the client sends a specific request to the server and receives a corresponding response. "Stateless" in this case means that each request is independent of previous ones and contains all information needed in order to be processed by the client or server.

Method	**Description**
POST	Creates a new resource
GET	Receives a representation of the specified resource
PUT	Updates or replaces the specified resource
DELETE	Deletes the specified resource

Table 2: Overview of HTTP request methods (CRUD operations).

4.1.2.1.1 Request methods

The HTTP protocol specifies different methods or verbs for requests that constitute actions that should be applied to the specified resource. The actual availability of a particular endpoint and the implementation details are defined by the server. Table 2 contains an overview of the verbs to create, receive, update, and delete resources (CRUD operations).

4.1.2.1.2 Request messages

HTTP request messages consist of four parts: a request line, several header lines, an empty line, and an optional body. The request line defines the HTTP request method, the endpoint, and the version of the HTTP protocol. The header contains the host and a variety of optional header fields such as the accepted format or the used language. The body contains the data that should be transmitted.
Listing 2 depicts an example of a GET request.

Listing 2: Request line and header of an HTTP GET request for the genetrail math api. In particular, the server is asked to calculate the square root of 16.

```
GET /api/math/sqrt/16 HTTP/1.1
Host: genetrail.bioinf.uni-sb.de
Accept: application/json
```

4.1.2.1.3 Response messages

HTTP response messages have a similar structure as requests. The first line of the header is a status line that contains the version of the protocol, a status code, and a status message that indicates if the request was successful. The response message then contains several header fields, followed by an empty line and an optional body. The body contains the representation of the requested resource.

Listing 3: Header and body of the HTTP response for the request in Listing 2. The body contains the response in json format.

```
HTTP/1.1 200 OK
Content-Type: application/json; charset=utf-8

{
    "result" : 4
}
```

4.1.2.2 *Hypertext Transfer Protocol Secure (HTTPS)*

The Hypertext Transfer Protocol Secure (HTTPS) [435] is an extension of the standard HTTP, where the communication is encrypted using Transport Layer Security (TLS) [539].

4.1.3 *Representational State Transfer (REST) and RESTful APIs*

Representational State Transfer (REST) is an architectural style for web applications that was developed as part of the doctoral dissertation of Roy Fielding [150]. The goal of REST is to provide guidelines for the implementation of web services that, amongst others, ensure good performance, scalability, and flexibility. To this end, Fielding defines six properties web services and APIs must fulfill in order to be considered RESTful.
Roy Fielding was also part of the development teams for the URI standard [50] as well as the HTTP/1.1 protocol [149]. Although REST constitutes a general design pattern of web interfaces, which is independent of the used protocol and implementation details, it shares common ideas and design principles with both standards. Hence, an

implementation of RESTful APIs using those standard web technologies has the advantage that they automatically fulfill some of the required properties.

4.1.3.1 *Client-server*

The first property of RESTful interfaces is that they have to implement the client-server model (cf. Figure 42). In this model, the user interface is separated from the data storage and implementation details of the server. In order to communicate, the client sends a request for a resource to the server. The server then processes this request and responds with a representation of the requested resource.

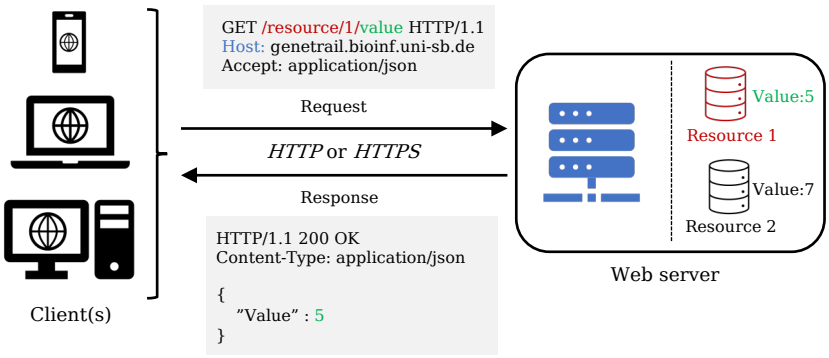


Figure 42: Example of a client-server architecture. Clients can send an HTTP request to query a specific resource on the server and receive a representation of this resource as response.

4.1.3.2 *Stateless*

The communication between client and server is additionally required to be stateless, which means that each message must contain all information to be processed by the client and the server, i.e., a message can not be split into multiple parts. This property ensures the reliability of the communications since all requests are self-contained and do not rely on previous requests. In case of a system failure or a significant time difference between two requests, this property also ensures that the communication can quickly be restored.

4.1.3.3 *Cacheable*

Response messages from the server are required to indicate in the header if they are cacheable (cf. Listing 3). In case of cacheable response messages, caching layers on the server, a proxy, or the client are allowed to save the response for later identical requests. This property of RESTful services has the potential to reduce network usage and improve performance.

Listing 4: Header and body of the HTTP response for the request in Listing 2. The header line colored in red indicates that the response can be cached and is valid for 1200 seconds.

```
HTTP/1.1 200 OK
Content-Type: application/json; charset=utf-8
Cache-Control: public, max-age=1200
{
    "result" : 4
}
```

4.1.3.4 *Uniform interface*

RESTful web services are required to provide a uniform interface with the following four properties:

1. All information managed by the web service is encoded as a resource with a unique identifier (e.g., a unique URI).
2. Different parts of the web server can then perform actions on a resource, and clients can receive a representation as a response that describes the complete state of that resource. The representation contains the data and metadata in the form of (key, value) pairs that are sufficient to create, query, update, or delete the addressed resource.
3. All messages sent between client and server have to be self-descriptive, which means they contain all information needed to understand and process the request or response.
4. Clients should be able to navigate the REST interface using hyperlinks provided by the server. This concept is called hypermedia.

Layered system

RESTful systems should also be implemented using a layered architecture (cf. Figure 43). Here, different functional components are encapsulated into individual levels that generally communicate through well-defined interfaces. This makes it possible to exchange implementation details without affecting the communication between different levels of the architecture or between client and server.

Code on demand (optional)

The last (optional) property of RESTful web services is that source code (e.g., scripts or applets) can be transmitted to the client. This property can reduce the computational burden of the server and the network load.

4.2 ARCHITECTURE

Author contributions

The first version of the Graviton framework was developed for the NetworkTrail web service and implemented by Oliver Müller, Daniel Stöckel, and me [509]. The current multi-layer architecture described in this chapter was initially developed for the GeneTrail2 web service, mainly implemented by Daniel Stöckel and me [510]. Since then, the framework has been continuously maintained, extended, and improved. Currently, it forms the basis for seven web services or databases and for which I was one of the leading developers. The following chapters of this thesis contain further "Author contribution" boxes with more specific information on each web service, and the complete list of contributors can be found in the author list of the respective publications [33, 177, 474, 475, 509, 510, 530, 533].

For the implementation of all our web services, we created a common framework, called Graviton, that was designed using the REST design principles as guideline (cf. Section 4.1.3). Accordingly, our framework is built using a layered architecture (cf. Figure 43) with seven distinct layers that are divided into three modules: (1) The front end with two layers on the client-side, (2) the back end with four layers on the server-side, and (3) a database layer. In the following sections, the individual components are described from top to bottom.

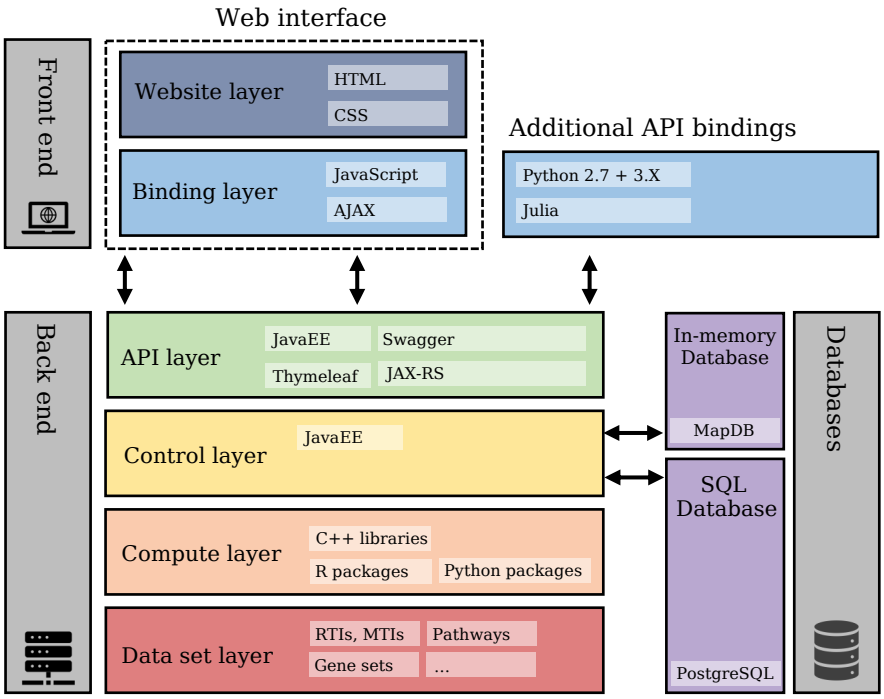


Figure 43: Different layers of the Graviton architecture.

4.2.1 *Front end*

The front end of our framework has two basic modes of operation. It can be accessed via an interactive web interface or programmed using an API.

4.2.1.1 *Web interface*

The web interface is the default mode of operation for all our web services. Here, users are guided through interactive interfaces, which allow them to upload data sets, select the parameters for their analysis, and visualize the results. All web interfaces are implemented using HTML [427, 553] and CSS [66, 552] with a custom layout created with the Bootstrap toolkit [523]. All computations on the client-side are implemented using JavaScript [132]. For the visualization of results, we also rely on several third-party JavaScript libraries: DataTables [316] to create interactive tables and Highcharts [24], D3 [67], and Plotly [413] for interactive plots. The individual web sites use JQuery [161] or AJAX [172] bindings to communicate and interact with the API layer of our web server. A screenshot of the web interface is depicted in Figure 44.

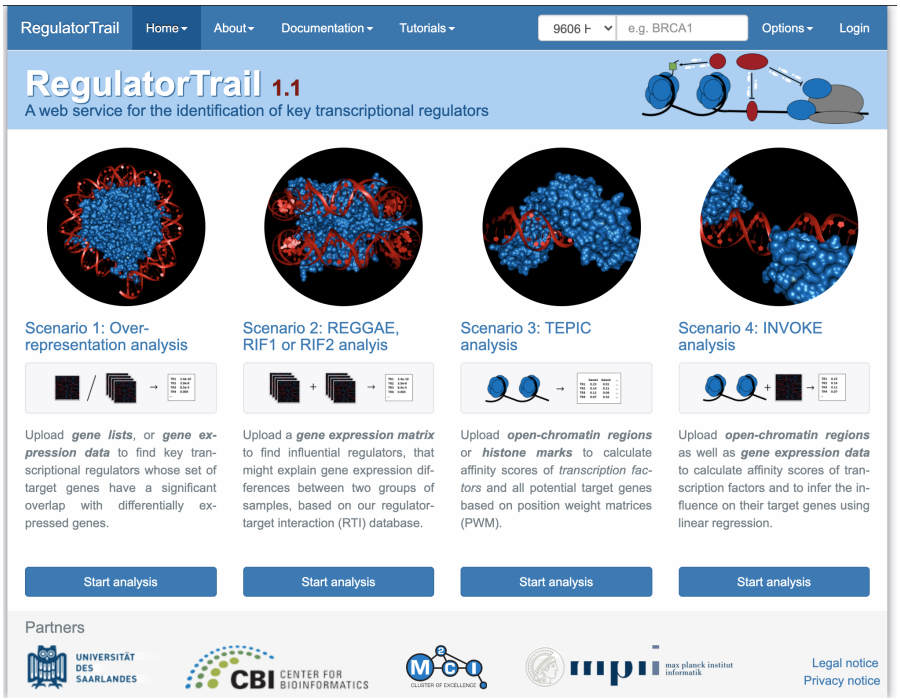


Figure 44: Screenshot of the RegulatorTrail start page.

4.2.1.2 *API bindings*

Apart from the web interface, users can also directly interact with our API. This is possible for all programming languages that support HTTP or HTTPS requests (cf. Section 4.1.2.1). Through interactions with the API, the web service can be controlled remotely. Hence, this feature allows to easily integrate our web services into third-party pipelines, such as Galaxy [181] or Taverna [583] workflows. In order to facilitate this process, we have already created Python and Julia packages that offer methods for the communication with our web services. These can be applied to run analyses on our web service, query corresponding results within the source code of these programming languages, and process the associated results. Additionally, we created a comprehensive documentation for all API endpoints using Swagger [499], which helps to create and test custom bindings.

4.2.2 *Back end*

The back end of our framework consists of four hierarchical layers with different levels of abstraction that are executed on the server-side.

4.2.2.1 *API layer*

The first layer in the back end is an application programming interface (API) that manages the client-server communication on the server side. To this end, it defines a set of operations that can be performed by our framework, such as uploading a microarray data set, conducting a group comparison, or creating a visualization of a particular result. Independent of the actual task, each operation is implemented as a specific API endpoint that processes the request, verifies all parameters, and then performs a specific action on the server. In our framework, we use two distinct types of endpoints: Jakarta Servlets [160] for web sites and Jakarta RESTful Web Services (JAX-RS) [159] based endpoints for the remaining API operations.
A Jakarta Servlet [160] is a Java class that processes HTTP requests for a specific URI. Based on this request and the submitted parameters, the servlet generates dynamic HTML pages that are then sent to the client. In our case, the HTML pages are generated using the Thymeleaf [524] template engine.
All remaining API endpoints are implemented as Java classes that use the JAX-RS specifications and annotations to handle specific API operations. These classes process HTTP-specific requests, verify the parameters and return the associated HTTP response in the requested format, e.g., JSON [74] or plain text. Here, the actual functionality is encapsulated in Java classes in the control layer (cf. Section 4.2.2.2) that provide a well-defined interface for the communication between

API and programs in the compute layer. This abstraction ensures that the API and the implementation of a particular task are independent and strictly separated. Hence, both layers can be modified or even exchanged without affecting the other one, which makes them easy to maintain.

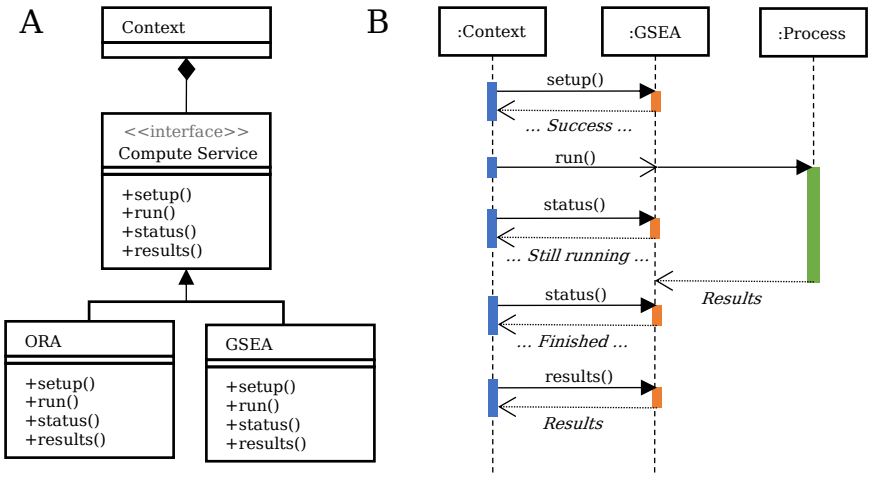


Figure 45: UML diagrams for an example compute service (GSEA). (A) UML class diagram depicting the strategy pattern implemented for compute services (GSEA and ORA). (B) UML sequence diagram for an asynchronous execution of a GSEA. From a specific context, i.e., our API, the GSEA compute service can be set up, started, and the results can be queried.

4.2.2.2 *Control layer*

The control layer of our framework consists of a collection of Java classes, hereafter called compute services, that encapsulate a particular task provided by any of our web services. These jobs can either be an individual program, several consecutive steps of a pipeline, or in some cases complete workflows, such as all steps of an enrichment analysis (cf. Section 3.6). Each of these Java classes provides a uniform interface that is shared amongst all compute services (Strategy design pattern [168]; cf. Figure 45). This interface allows the respective job to be set up and controlled by other classes in the control layer or associated API endpoints. The actual programs are executed as an asynchronous process that is monitored by the corresponding compute service. The current status of this process (e.g., the progress), exit codes, and all associated results can then be queried via the interface of this compute service. The execution as an asynchronous process ensures that clients that interact with an API endpoint receive immediate feedback about the status of a job and do not have to wait for the computation to finish, which frees up computational resources that would be otherwise wasted. Each compute service ad-

ditionally interacts with our SQL database to log the metadata of all computations, e.g., inputs, parameters, and provided outputs.

4.2.2.3 *Compute layer*

The compute layer of our framework contains the actual implementation of all tasks performed by our web services. Here, each task is implemented as an individual program that is controlled by the compute services in the control layer. In the following paragraphs, we give a brief overview of how the different programs are implemented.

GeneTrail2 C++ library

Most programs are implemented in a C++ library we initially created for the GeneTrail2 web service. This library consists of different modules:

1. The core module contains classes that implement hypothesis tests, statistical measures, parsers for various input formats, and data structures to represent the different data types.
2. The enrichment module builds upon the core functionality and provides the implementation of all supported enrichment algorithms (cf. Chapter 5).
3. The regulation module extends the core functionality with implementations of many different approaches to analyze the influence of transcriptional regulators. (cf. Chapter 8).

The library was created using a custom CMake build system [335, 336] and depends on several third-party libraries: (1) The Eigen library for matrix and vector operations [241], (2) the Boost library, e.g., for file handling and mathematical distributions [469], (3) OpenMP for parallel computation [108], and (4) GoogleTest for unit testing [312].

C++ libraries for integer linear programs (ILPs)

Our web services also offer a variety of integer linear programs (ILPs): (1) Subgraph ILP (cf. Section 3.7.1), (2) Transitivity Clustering ILP (cf. Section 5.5.2.2), (3) Maximum targetome coverage ILP (cf. Section 6.3.2.1). We implemented all of them as C++ programs using the ILOG CPLEX Optimization Studio [330].

TEPIC

For the `RegulatorTrail`, we additionally use the `TEPIC` framework developed by Schmidt et al. [472] to calculate transcription factor affinities based on position-weight matrices and to infer the influence of transcription factors using linear regression (cf. Chapter 8).

R and Python scripts

Some of the tasks are also implemented as Python or R scripts, such as hierarchical clustering or dimension reduction.

4.2.2.4 *Data set layer*

The last layer in the back end represents our large collection of third-party data sets (cf. Section 3.2) that are the basis for all our workflows. All contained data sets are carefully selected, curated, and sanitized. This ensures that the data sets of the different workflows are compatible with each other.

4.2.3 *Database layer*

Our framework also uses two databases: One in-memory database that handles identifier mapping and one SQL database that manages the data sets and results for all analyses.

4.2.3.1 *In-memory database*

The in-memory database is implemented using MapDB [280]. It contains mapping tables for the most common identifier types for genes, proteins, and miRNAs, as well as conversions between them. All mapping tables are directly stored in the main memory, which eliminates disk access and improves the response time for requests. How this database is used to sanitize all uploaded data sets is described in Section 4.3.2.

4.2.3.2 *SQL database*

Our framework additionally employs a traditional SQL database (i.e., PostgreSQL) to manage the analyses and results of all users. For each data set, this databases documents metadata such as the type of the data set, the format, the organism, or the identifier type. This information is then analyzed to determine which analyses can be performed using this file (cf. Section 4.3.1). For all analyses, the configured parameters and the produced results (including intermediate ones), the status codes (e.g., error messages), and timestamps are saved. This ensures transparency and reproducibility of all analyses and allows to verify the parameters of all computations.

4.3 GENERAL FUNCTIONALITY

Our framework also provides general functions for the validation and management of the different input files that are shared amongst all our tools.

4.3.1 *File type validation and sanity checks*

Graviton provides several routines that automatically analyze and validate all input files uploaded by a user. For each file, sanity checks are performed that verify the format and that determine and normalize the encodings. These checks ensure that the uploaded files are valid and can be properly analyzed in subsequent steps.
Additionally, these routines extract meta information from the data set to select suitable parameters for the analysis automatically. The extracted metadata contains properties like the dimensions of the data, gene or protein identifier, organism, as well as numerical properties. This information is then examined to check if a specific analysis step can or needs to be performed. For example, if the uploaded file is a matrix with different sample groups, a group comparison is performed that is omitted for different file types (e.g., score list or identifier list). The dimensions of the data and the numerical properties additionally define which statistical tests can be applied, e.g., if negative values are present, all methods that include logarithm cannot be applied.

4.3.2 *Identifier mapping*

Depending on the experimental technique applied to generate a particular data set, the repository where a data set is deposited, or the age of the experiment, different identifiers (ids) are used to denote a certain molecular feature. On the one hand, these ids distinguish different aspects or states of a feature (cf. Figure 46). For example, a gene may be referred to by its gene symbol (e.g., PDCD1), the id of a particular transcript (e.g., ENST00000334409), or the functional gene product (e.g., protein Q15116). On the other hand, data sources like large repositories or large research projects often employ other standards to define a particular feature. For example, genome annotations for the human genome are, amongst others, provided by NCBI [420], Ensembl [595], or the HAVANA project [203]. All adopt different guidelines to define genes or transcripts, and as a consequence, all projects provide custom identifier types for each feature. Although there already have been efforts to define a common standard for human and mouse genomes (CCDS [410, 419]), there is still a large number of identifier types that are regularly used.

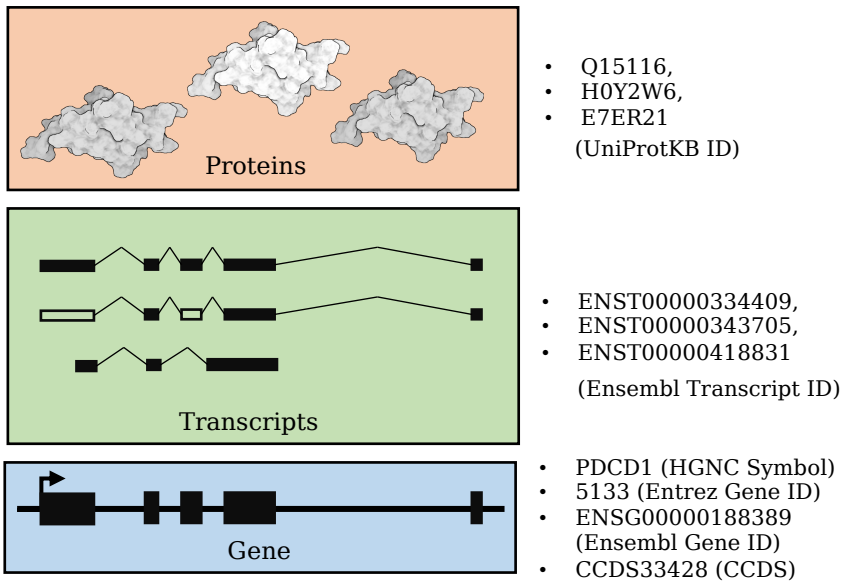


Figure 46: Overview of different identifier types for the Gene PDCD1. This figure was created using icons from BioRender.com.

Hence, for most computational analyses, data sets need to be carefully curated and sanitized before they can be combined with other data sources. For the most part, the ids of different annotations are compatible, but there are exceptions where features are only part of one annotation. Moreover, the different databases are regularly updated, and features might be replaced or renamed. In the following, we briefly describe the different procedures that automatically detect the provided identifier type or convert them into a uniform representation that is compatible with all data sets in collection. For this purpose, we utilize an in-memory database (cf. Section 4.2.3.1) that contains reference lists and mapping tables for gene, protein, SNP, and miRNA identifiers from well-curated databases (cf. Section 3.2.1). The in-memory database is used in a three-step mapping approach. In the first step, the reference lists are examined to detect the organism and identifier type of each uploaded data set. To this end, we check for each identifier in the data set in which reference sets it is contained. By default, we assign the organism and identifier type using majority voting. Alternatively, users can select, verify, or change this information manually.
In a second step, the identifiers of all uploaded data sets are mapped from the assigned id type to "Official Gene Symbols" in order to be compatible with all internal databases of our framework. To this end, we process all ids individually. First, we verify if a considered id is contained in the mapping table, i.e., whether it is a valid identifier. In this case, the identifier is directly mapped. In case it is not contained in the mapping table, we check if the id is a known alias or belongs to a previous version of the annotation, if corresponding mappings

exist, and if we can map the id. In instances where no valid mapping is found, the respective entry is removed from the data set.
During this mapping procedure, we could potentially introduce duplicates that need to be removed before the file is further processed. To this end, our framework provides four different methods that can be applied to aggregate the values of duplicate entries, i.e., mean, median, max, or min.

5

GENETRAIL WEB SERVICE

Author contributions

In this chapter, we describe GeneTrail, a web service for enrichment and network analysis of molecular high-throughput profiles. The original version of this tool was developed by Christina Backes, Andreas Keller, and Hans-Peter Lenhof [29, 257]. Here, we describe the second and third versions of the web service, for which I was one of the leading developers [176, 177, 509, 510]. This chapter consists of several parts that contain further "Author contribution" boxes with specific information on each part. The complete list of contributors can be found in the author list of the respective publications [29, 176, 177, 257, 509, 510].

In current biomedical research, experimental high-throughput techniques, such as high-throughput sequencing or microarray experiments, are regularly applied to profile cells on a large scale, both on a bulk and single-cell level (see, e.g., [101, 207]). Independent of the used protocol, the resulting data sets are usually high-dimensional and often prone to technical noise [35, 294, 547].
Hence, powerful computational analysis methods are required that facilitate the exploration of these data sets and that help researchers to gain novel insights into the underlying biology.
In this context, one important task is the identification and analysis of deregulated biological processes and, in particular, processes that show significant differences between two sample groups, e.g., disease vs. control. Amongst the most important methods for this purpose are enrichment or network analysis algorithms as discussed in Sections 3.6 and 3.7.
Due to their popularity, a variety of tools have been published that implement respective approaches, both as stand-alone tools and web services. In general, they can be distinguished based on the supported algorithms, data types, databases, or organisms (cf. Figure 47). For example, both LOLAweb [375] and miEAA [30] focus on specific data types. LOLAweb analyzes (epi-)genomic regions, and miEAA performs enrichment analysis for miRNAs. Further tools, such as DAVID [227], Enrichr [284], GSEA-P [513], or RegulatorySnapshots [183], focus solely on enrichment analysis, while others, like Babelomics [16], iPEAP [515], Paintomics [120], RAMONA [466], and WebGestalt [305], additionally offer network analysis procedures.

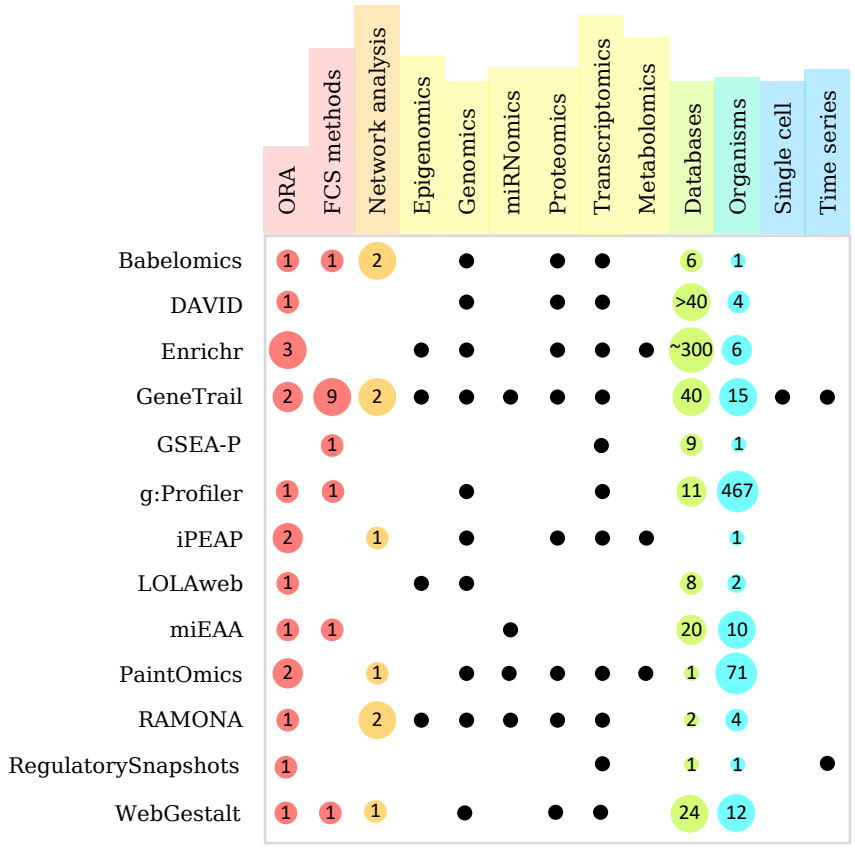


Figure 47: Comparison between GeneTrail and other enrichment or network analysis toolboxes.

In this chapter, we describe GeneTrail, a web service for integrative analysis of (epi-)genomic, miRNomic, proteomic, and transcriptomic data sets [29, 176, 177, 257, 509, 510]. The rich functionality of our web service can be applied to study deregulated biological processes in bulk, time-series, and single-cell data sets. To this end, we created a powerful framework of enrichment and network analysis methods that can be used to explore our comprehensive collection of predefined biological categories for 15 different organisms (cf. Appendix G). GeneTrail was built using our Graviton web framework described in Section 4.2. All algorithms are implemented as highly optimized C++ programs that are also available as stand-alone applications on GitHub.

In the following sections, we describe the different application areas of our web service and demonstrate some of the key aspects based on real-world data sets.

5.1 STANDARD ENRICHMENT ANALYSIS

Author contributions

The standard enrichment analysis functionality of GeneTrail is a complete redesign and major extension of the original GeneTrail web service [29, 257]. It was introduced with the second version of GeneTrail and mainly developed by Daniel Stöckel, Hans-Peter Lenhof, and me. All contributors are listed in the author section of our publication "Multi-omics enrichment analysis using the GeneTrail2 web service" [510].

The core of the GeneTrail web server are the standard enrichment analysis workflows for genomic, proteomic, transcriptomic, and miRnomic data sets. For each of these omics types, users can upload molecular measurements in various formats. After file upload, they are guided through an intuitive step-by-step interface, which helps them to build customized pipelines and to choose appropriate parameters for the analysis. Here, automatic routines, like the ones described in Chapter 4, analyze the uploaded data and preselect suitable algorithms and parameter combinations.
In the following paragraphs, we describe the input formats, the general workflow, and the different visualization types for enriched biological categories. An overview of all possible workflows is depicted in Figure 48.

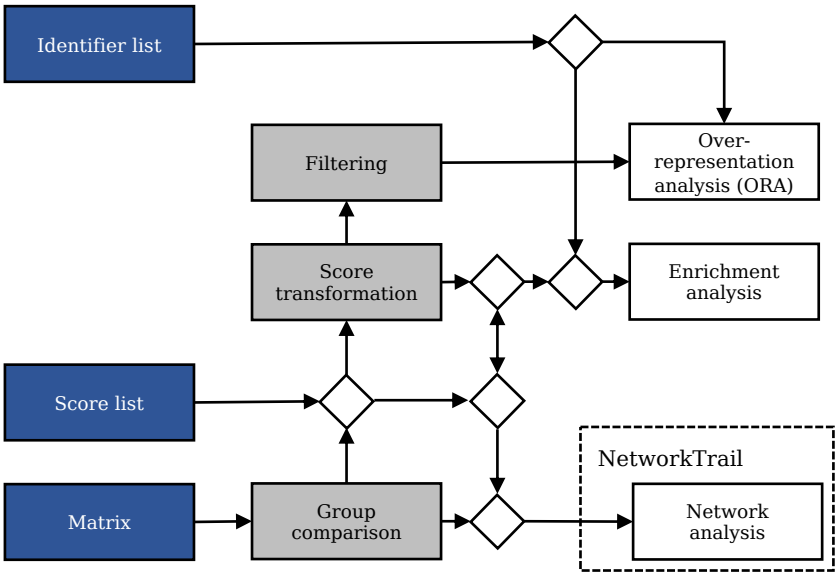


Figure 48: Overview of the GeneTrail standard workflow. The different input types are marked in blue. Intermediate processing steps are shown in grey, and the different analysis types are depicted in white. The arrows mark all possible workflows provided by our framework. This figure was adapted from Stoeckel et al. [510].

5.1.1 *Input data*

The input data for the standard enrichment analysis workflow can be uploaded in three different ways: (1) Feature lists, (2) score lists, and (3) feature matrices. All of them can be uploaded as whitespace-separated files that contain one line per feature.

All uploaded files are automatically sanitized, and the identifier for all features are mapped to a uniform representation (e.g., HGNC symbols for genes, cf. Section 4.3.2).

5.1.1.1 *Feature lists*

Feature lists can either contain unordered sets of molecular features that should be used in an over-representations analysis or a sorted list of features for which we can also apply non-parametric enrichment analysis methods, such as the gene set enrichment analysis (cf. Section 3.6.3.1).

5.1.1.2 *Score lists*

In addition to the feature name, score lists additionally contain a weight per feature that should reflect its importance, such as a score from a group comparison. This allows users to apply scoring methods that might not be supported by our framework.

5.1.1.3 *Feature matrices*

Feature matrices contain different sample measurements for each feature. Hence, matrices are required to contain a header field that contains a unique identifier for each sample. An example is shown in Listing 5.

Listing 5: Example of a feature matrix with measurements for three genes in the samples.

```
Sample1   Sample2   Sample3
Gene1   0.0   0.1  3.4
Gene2   4.0   4.1  3.9
Gene3   5.7   6.5  1.2
```

5.1.2 *Workflow*

Similar to the "general modular framework for gene set enrichment analysis" proposed by Ackermann and Strimmer [2] (cf. Section 3.6), each enrichment analysis workflow in GeneTrail is divided into different processing steps:

1. Group comparison (or feature-level statistic)
2. Score transformation
3. Enrichment analysis (Set-level statistic + P-value strategy)
4. Multiple-testing correction

For each step, our framework offers different methods that can be combined to create customized pipelines that are tailored to the analyzed data set. In the following paragraphs, we give a brief overview of the individual steps and the provided methods.

5.1.2.1 *Group comparison*

While feature lists and score lists can directly be used for enrichment analysis, matrices need to be processed first. To this end, GeneTrail provides a variety of methods for the comparison of two sample groups within this matrix.

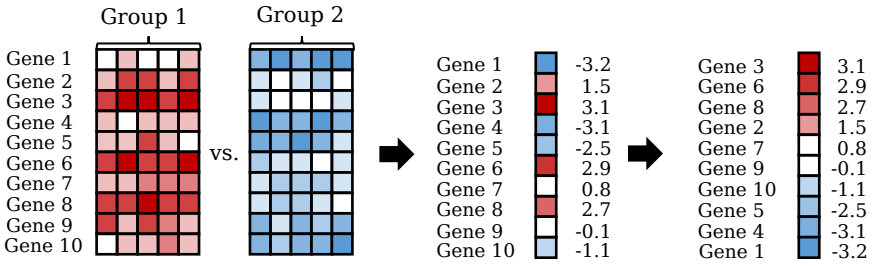


Figure 49: Overview of the group comparison step of GeneTrail. Here, a test statistic is applied to calculate scores that distinguish the two sample groups.

In addition to standard (log2-)fold-changes (cf. Section 3.4), these can be divided into three groups that are presented in the following.

Parametric tests

GeneTrail offers various parametric tests that can be applied to compare two sample groups. For microarray analysis, we implemented several methods assuming the samples are drawn from a normal distribution, in particular, members of the t-test family, such as the dependent t-test for paired samples [252], the Welch's t-test [572], and

the Shrinkage t-test [398] (cf. Section 3.4). For the analysis of count data, e.g., from RNA-Seq experiments, GeneTrail also provides specialized methods, like DESeq [20], edgeR [443], and RUVSeq [441].

Non-parametric tests

Besides parametric tests that assume a specific underlying distribution, our web service also offers non-parametric alternatives that can be applied to all data sets without any restrictions on the distribution of the data, e.g., the Wilcoxon rank-sum test [440] and the Wilcoxon matched-pairs signed-ranks test [440].

Correlation coefficients

In addition to typical group comparison methods, our framework also provides correlation coefficients that can be applied for this purpose. To this end, we a calculate the correlation between the molecular measurements of each feature and a binary vector that indicates the group assignment of the respective samples, e.g., disease and control.

5.1.2.2 *Score transformation (optional)*

The next step in the standard enrichment analysis workflow is the transformation of uploaded or computed score lists (cf. Figure 50). For this purpose, our framework offers five different strategies: (1) no transformation, (2) square root, (3) logarithm, (4) squared, or (5) absolute values.

In the context of enrichment analysis, the square root and the logarithm can help to reduce the effect of outliers, while squared values and absolute values can facilitate the detection of biological categories that contain both up- and down-regulated features.

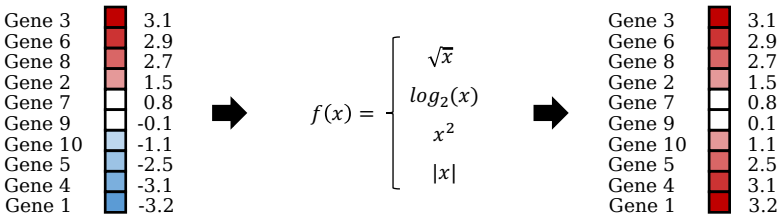


Figure 50: Overview of the score transformation step in the GeneTrail workflow.

5.1.2.3 *Enrichment analysis*

The next step in our workflow is the actual enrichment analysis. Here, we distinguish between two distinct strategies: Over-representation analysis (ORA, cf. Section 3.6.2) or functional class scoring (FCS, cf. Section 3.6.3). An overview of both strategies is depicted in Figure 51.

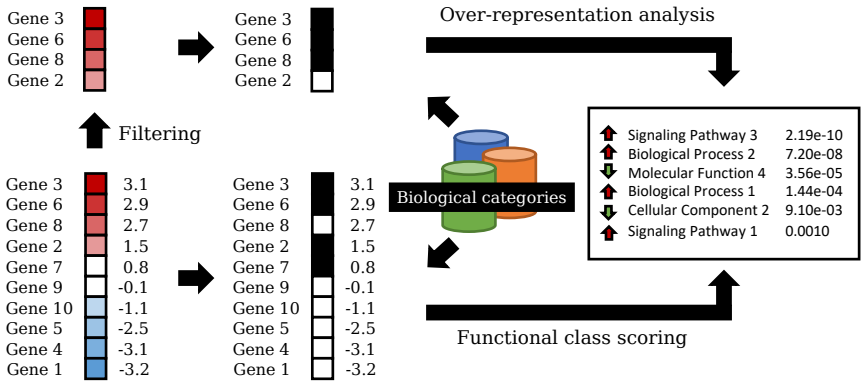


Figure 51: Overview of the enrichment analysis step in the GeneTrail workflow. The input score list can either be used directly in an enrichment analysis, or filtered and used in an over-representation analysis (ORA).

5.1.2.4 *Over-representation analysis*

The input for an ORA is a small feature set, such as the most up-regulated genes. The feature set can either be created by filtering a score list or directly uploaded by the user. Our web service then tests if the feature set contains significantly more (or less) elements of a biological category than expected based on a reference set (cf. Section 3.6.2).

5.1.2.5 *Functional class scoring*

Unlike ORA, the FCS methods utilize all measured features to detect deregulated biological categories. To this end, they compare all features in the input list that belong to a certain category with all others. GeneTrail provides several methods that employ this strategy (cf. Section 3.6.3). In general, they can be divided into three classes: parametric tests, non-parametric tests, and averaging methods.
As parametric tests, we implemented the one-sample t-test [603] and the Welch t-test [572]. Both tests use the actual scores to compare the distribution of the features in the category to all other features. To this end, they assume that both groups are drawn from Gaussian distributions.

In contrast, non-parametric tests make no assumptions about the distribution of the scores and compare the two groups based on their ranks in the sorted score list. For this class, we implemented three methods: the Wilcoxon rank-sum test [440] and both the weighted [514] and the unweighted [259] version of the gene set enrichment analysis (GSEA, cf. Section 3.6.3.1).
Moreover, our framework also offers several averaging methods (cf. Section 3.6.3.3). These approaches calculate a sample statistic for all elements in a biological category and then assess the significance of this score via a permutation test. GeneTrail offers four such sample statistics: mean, median, sum, and max-mean [136].
For both ORA and FCS approaches, our framework provides different strategies to calculate p-values. For all methods, p-values can be estimated using either the sample- or feature-based strategy proposed by Ackermann and Strimmer ([2], cf. Section 3.6.1). For the sample-based strategy, all p-values are calculated via permutation tests. For the feature-based strategy, p-values can also be derived from the approximated probability distribution of the test statistic or even calculated exactly (cf. Table 3).

		Sample-based permutation	Feature-based permutation	Approximated (distribution)	Exact
	ORA	x	x		x
Parametric	One sample t-test	x	x	x	
	Welch t-test	x	x	x	
Non-parametric	Unweighted GSEA	x	x		x
	Weighted GSEA	x	x		
	Wilcoxon rank-sum test	x	x	x	
Averaging	Max-mean	x	x		
	Mean	x	x		
	Median	x	x		
	Sum	x	x		

Table 3: Overview of p-values strategies available for the enrichment methods implemented in GeneTrail.

5.1.2.6 *Multiple testing corrections*

In general, the enrichment analysis workflow is applied to multiple biological categories simultaneously. Since we use a predefined significance level to assess if a biological category is significantly enriched or depleted, we run into the multiple testing problem, as described in Section 3.3.2. Hence, GeneTrail provides a variety of methods that either control the family-wise error rate (e.g., Bonferroni [63, 64]) or the false discovery rate (e.g., Benjamini-Hochberg [44] or Benjamini-Yekutieli [45]).

5.1.3 *Visualization of results*

On top of the rich functionality described in the previous paragraphs, GeneTrail also creates different visualization of the computed results. These range from a general overview of enriched biological categories to an in-depth characterization of individual ones, as well as relationships between them. Examples for the different types of visualizations can be found in Figure 53.

5.1.4 *Integrative analysis*

Besides the standard workflow that can be applied to examine the results of one analysis, our web service additionally offers functionality to integrate and combine the results of different analyses. To this end, users can select multiple enrichment results and view them side-by-side as a single table. In general, our web service offers two options to combine enrichment results, the intersection or the union of all significantly enriched or depleted categories. This functionality can be used to quickly compare similarities and differences between different analyses, or to study how pathway activity patterns change across multiple data sets. An example for this is depicted in Figure 52.

Name	Patient 1	Patient 2	Patient 3
adaptive immune response	6.70e-14	1.0000	1.08e-19
negative regulation of immune system process	2.32e-10	1.0000	7.45e-14
positive regulation of immune response	2.46e-10	2.18e-6	8.57e-30

Figure 52: Comparison between GSEA results of the tumors of three different patients. This table was created using a hepatocellular carcinoma data set (GSE64041; cf. Section E.1).

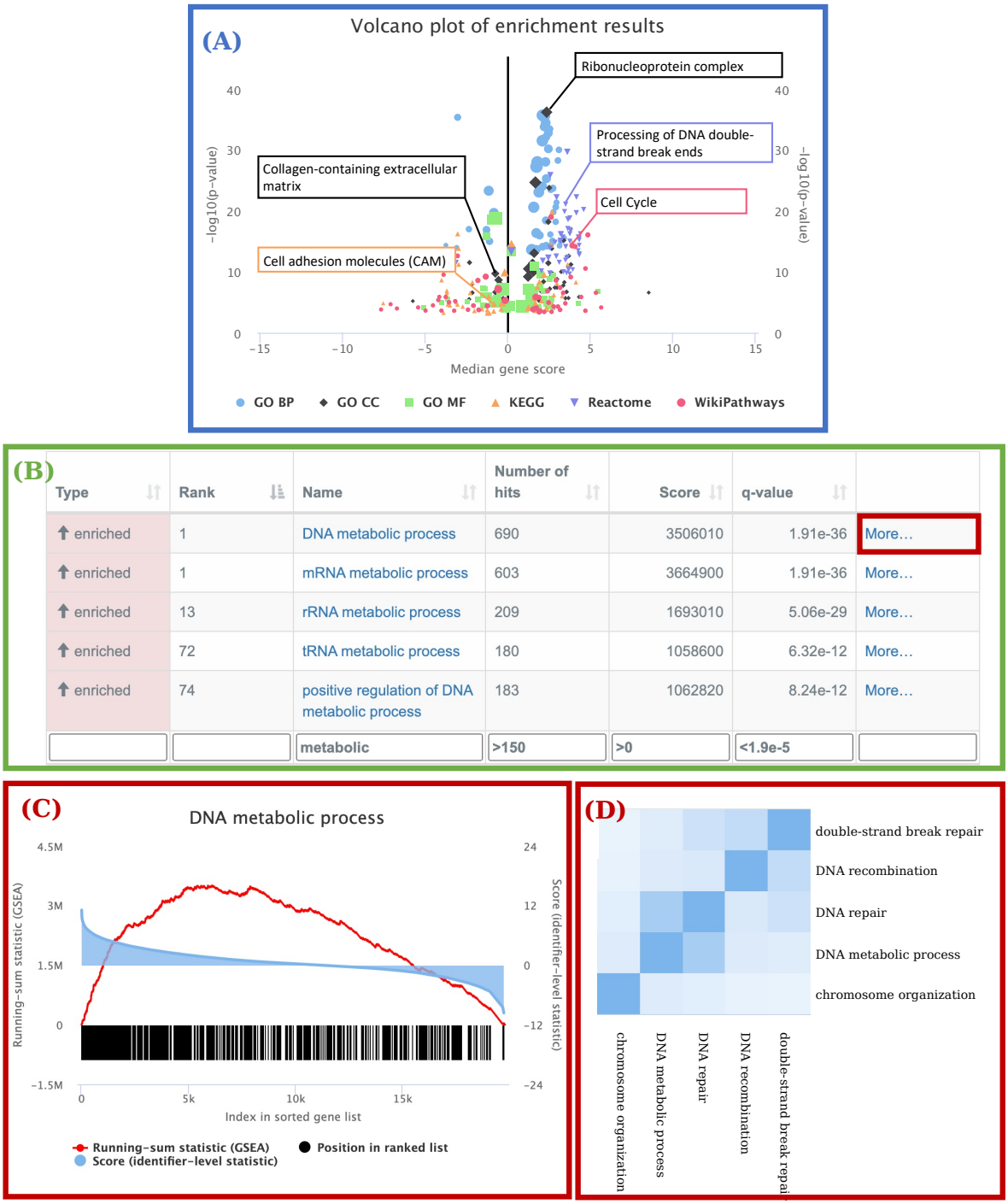


Figure 53: Overview of the different types of visualizations available for enrichment results. (A) Volcano plot where each category is represented as a point in the scatter plot. (B) Table depicting significantly enriched/depleted categories for a specific database. (C) GSEA running-sum plot (red), density plot of gene scores (blue), and positions of genes in ranked list (black) for category "DNA metabolic process". (D) Heatmap depicting the gene set overlap between the genes of "DNA metabolic process". The plots were created using a hepatocellular carcinoma data set (GSE64041; cf. Section E.1).

5.2 NETWORK ANALYSIS

Author contributions

The network analysis functionality of our web service was originally developed for the NetworkTrail web service that was implemented by Daniel Stöckel, Oliver Müller, and me [509]. With the introduction of the layered architecture of our framework and the corresponding modular web interface (cf. Section 4.2, [510]), we migrated the complete functionality into the updated framework. This was mostly done by Daniel Stöckel and me. The complete list of contributors can be found in the author section of the respective publications [509, 510].

In addition to the classical enrichment analysis workflows described in the previous section, our framework also offers network analysis algorithms. Instead of biological categories that represent feature sets with a particular annotation, network analysis methods examine the topology of interaction networks to identify deregulated subgraphs or even signaling cascades (cf. Section 3.7). These methods cannot only be used to study deregulated biological processes, but also to identify important elements within them, such as the root.
The workflows described in this section were initially developed for the NetworkTrail web service [509], but are also part of the GeneTrail web service [177, 510]. An overview of the different processing steps is shown in Figure 54.

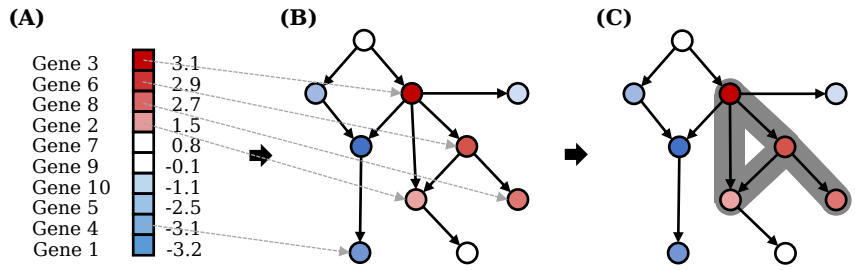


Figure 54: Overview of the NetworkTrail workflow. (A) The input for the network analysis workflows is a score list (cf. Section 5.1.1), where each score indicates the importance of a specific feature. (B) The scores in this list are used as weights for corresponding vertices in the interaction network. (C) Both ILP and FiDePa analyze the topology of the resulting network to identify the most deregulated subgraphs.

5.2.1 *Workflow*

The input for all our network analysis algorithms is a score list, where each entry represents the importance of a specific feature, e.g., the degree of deregulation. This score list can either be uploaded by a user or be calculated based on a given gene expression matrix with two sample groups, e.g., disease vs. control. Since these processing steps (group comparison and score transformation) are identical to our enrichment analysis workflow, they are skipped in this section (cf. Figure 48).
In addition to the score list, the different algorithms also require an interaction network as input, where each vertex represents a feature and each edge an interaction between two features (cf. Section 3.2.3). The uploaded scores are then used as vertex weights in this graph.
Our framework offers two methods for the identification of the most deregulated subgraphs in the given network: FiDePa [258] and the ILP approach described in Section 3.7.1 [32]. FiDePa applies a dynamic programming approach to examine which paths (linear subgraphs) of a given length are significantly enriched in the score list. To this end, the Kolmogorov-Smirnov statistic (unweighted GSEA) is employed (cf. Section 3.6.3.1). In addition to linear subgraphs, the ILP is also able to assess branched subnetworks of predefined size (cf. Section 3.7.1). Both methods are usually applied to analyze different sizes of subgraphs, which are then combined for the final results.
Finally, our web service provides different ways to visualize the resulting graphs, i.e., either using the BiNA visual analytics tool [175] or directly in the browser using a custom Cytoscape.js plugin [162] (cf. Figure 55).

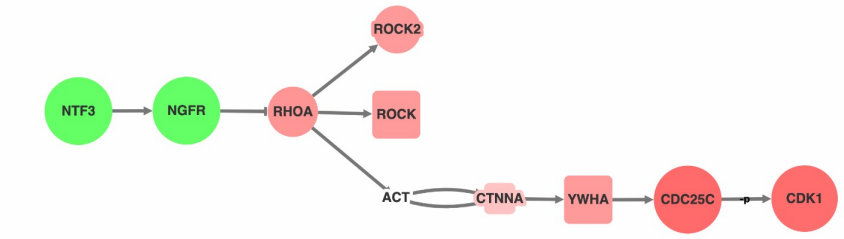


Figure 55: Example visualization of a subgraph with 10 nodes. The size and color indicate the degree of deregulation of each gene. The shape indicates if a node is a protein (round), protein family (rectangle), or protein complex (diamond; not shown). The plot was created using a hepatocellular carcinoma data set (GSE64041; cf. Section E.1).

5.3 EPIGENOMICS WORKFLOW

Author contributions

The epigenomics workflow was mainly developed by Nico Gerstner and Hans-Peter Lenhof. I assisted with the design and implemented the parallel version of the underlying ORA. The complete list of contributors can be found in the author section of the respective publications [177].

The chromatin structure in a cell is a crucial factor in the control of the transcription process (cf. Section 2.1.5). It defines if the DNA and in particular genes or gene regulatory regions (GRR) are accessible to DNA-binding proteins, such as transcription factors or the polymerase complex. The chromatin structure is induced and influenced by specific epigenetic modifications of the DNA or histone proteins, i.e., cytosine methylations or histone marks. Hence, the assessment of chromatin states in a biological sample provides important information about the gene expression and, subsequently, the activity of biological processes.

In the third version of GeneTrail, we introduced a specialized workflow for the integrative analysis of epigenetic modification patterns in different sample groups and their putative effect on biological processes and signaling pathways [176, 177].

The input for this workflow is a collection of epigenetic modification patterns for each analyzed sample, e.g., open-chromatin regions, histone marks, or DNA methylation patterns. These can be uploaded in different standardized file formats, which include "BED", "VCF", or "IDAT". All of them are described in Appendix F. Our web service then conducts the following processing steps to find genes that have a different chromatin state in two analyzed samples, e.g., disease vs. control, and subsequently which biological processes might be affected by these differences.

5.3.1 *Chromatin state assignment*

In the first step, GeneTrail investigates the chromatin structure of each sample to assign one of the following functional states to each gene: "active", "poised", "repressed", or "no information". To this end, our web service evaluates the chromatin modification patterns in the regulatory regions of all transcripts individually. This is done using manually curated rules that assess the combination of epigenetic marks in these regions to estimate the chromatin state of each transcript. The rules were either extracted from databases, i.e., Histome [262] and HHMD [610], or manually derived from literature (cf. Section 2.1.5.1).

In general, our data collection contains two types of rules. The first type are combinations of epigenetic modifications that are well-studied and clearly define a specific chromatin state. Examples for such rules are the occurrence of H3K4me1 and H3K27ac in the enhancer region or H3K4me3 and H3K27ac in the promoter region, which mark actively transcribed genes. The second type of rule are more general associations of a specific mark and the activity of a gene, such as the H3K4ac mark, which is enriched in the promoter of actively transcribed genes.
In order to combine both types of information, we employ an evaluation scheme that first assesses specific rules that provide strong evidence. If none of the specific rules are met, we check if the if the regulatory regions of a transcript predominantly contain marks that are associated with a given chromatin state.
Finally, we obtain one of the four chromatin states for each transcript. However, since we require gene-level information for most downstream applications, we have to aggregate the transcript information for each gene. To this end, we currently assign each gene to the functional state of its most active transcript.

5.3.2 *Identification of chromatin state transitions*

In a second analysis step, GeneTrail searches for genes with different functional states in two analyzed sample groups. Genes with the same behavior are then clustered to create "transition groups", e.g., all genes with a poised state in the first sample group and an active state in the second one. An overview of this procedure is shown in Figure 56.

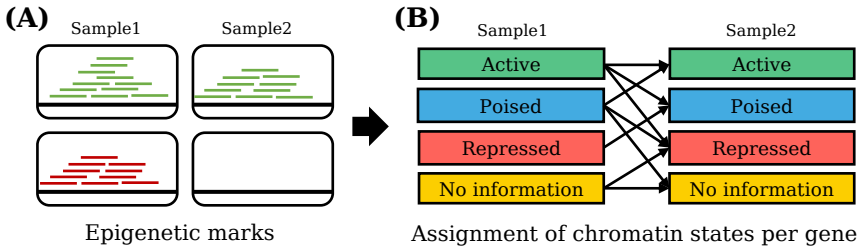


Figure 56: Overview of the GeneTrail epigenomics workflow (Part1). The combination of epigenetic modifications in the regulatory regions of each gene define one of four chromatin states for each sample (active, poised, repressed, no information). This figure was adapted from [177].

5.3.3 *Enrichment analysis*

In the final analysis step, we conduct over-representation analyses for the genes of each transition group (cf. Figure 57). This can be used to investigate which biological processes might be affected by the respective changes.

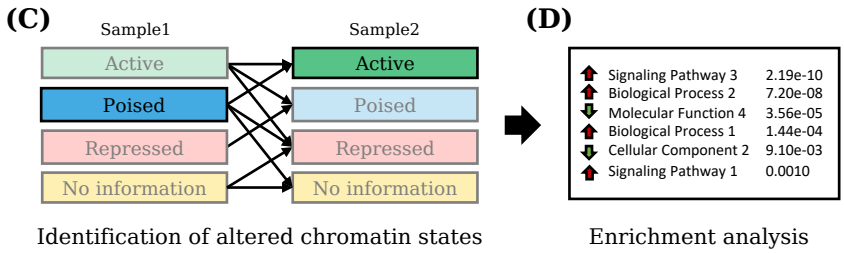


Figure 57: Overview of the GeneTrail epigenomics workflow (Part2). All genes that are assigned to a specific transition group (e.g., "poised" to "active") are used to conduct over-representation analyses. This figure was adapted from [177].

5.4 SINGLE-CELL ANALYSIS

Author contributions

The single-cell analysis functionality of GeneTrail was mainly designed and developed by Nico Gerstner, Hans-Peter Lenhof, and me. The front end was mainly implemented by Nico Gerstner, and the back end was implemented by Nico Gerstner and me. The complete list of contributors can be found in the author section of the respective publications [177].

Over the last years, single-cell RNA sequencing (scRNA-seq; cf. Section 3.1.1.3) technology has become a major focus in biomedical research. The measurements of expression levels on a single-cell basis instead of a cell mixture not only allow studyng biological systems at a previously unmet resolution but also to examine the different cell types in a sample. Hence, scRNA-seq has already successfully been used in many research projects. Amongst others, this technology was applied to catalog and study gene expression levels in many tissues and cell types of different organs in various organisms, e.g., mouse [101], fruit fly [456], or human [416]. In this context, researchers were even able to identify previously unknown cell populations [365]. Furthermore, single cell techniques have also been used to study expression differences between diseased and normal cells, e.g., for COVID-19 [581] or cancer [291, 422].

Modern scRNA-seq data sets can provide measurements for hundreds of thousands of single cells. Consequently, a manual inspection of the data is completely impossible, and highly efficient computational tools are required that help to process and explore the contained information.

For this purpose, we extended GeneTrail with powerful single-cell functionality [176, 177]. This toolbox provides several workflows for the analysis and exploration of huge scRNA-seq data sets. An overview of these workflows is provided in Figure 58.

Our web service was designed to assess active biological processes in individual cells and to examine functional differences between different cell groups, e.g., cell types or clusters. In order to efficiently handle the huge amount of data, we carefully revised, optimized, and parallelized all processing steps. We even developed a new index-based ORA approach that conducts an enrichment analysis in constant time. In the following, we first describe each processing step and then demonstrate our single-cell functionality using a COVID-19 data set (cf. Section 5.6.1).

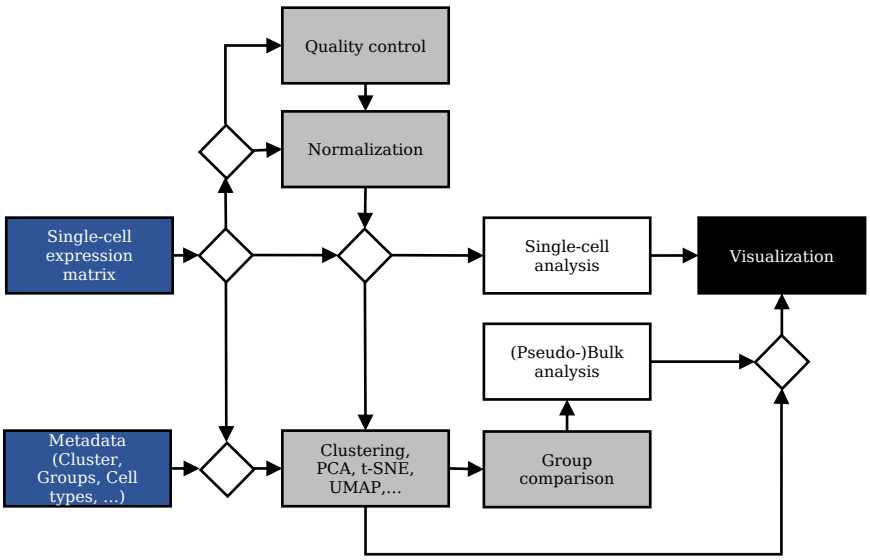


Figure 58: Overview of the GeneTrail single-cell analysis workflows. The different input types are marked in blue. Intermediate processing steps are shown in grey, and the different analysis types are depicted in white. The arrows mark all possible workflows provided by our framework.

5.4.1 *Preprocessing*

The input for our single-cell workflow is either a normalized gene expression matrix or a raw count matrix. These can be uploaded in our standard matrix format (cf. Section 5.1.1) or in special sparse formats that are described in Appendix F. While normalized expression matrices can directly be used for the analyses described in the subsequent sections, raw count matrices need to be processed first. This includes multiple quality control and normalization steps that are described in the following.

5.4.1.1 *Quality control and filtering*

In current research, a variety of experimental techniques can be used to isolate single cells and to generate associated expression profiles (cf. Section 3.1.1.3). Typically, the resulting data sets are represented as a count matrix, where each row represents expression measurements for one gene and each column a specific barcode. Here, we use the term barcode instead of cell because due to technical problems during the single-cell isolation, library preparation, or sequencing steps, reads assigned to one barcode may not necessarily originate from one single cell [320]. Instead, depending on the used protocol, one barcode might mistakenly denote measurements of two or even more cells that were not properly separated, i.e., doublet or multiplets. In

some protocol, we could also obtain no measurements at all, e.g., if barcodes are assigned to empty droplets or wells [320]. Furthermore, during the single-cell isolation step, it is also possible that cells are damaged and that only a part of the RNA can be measured.
In all cases, the values of these barcodes might influence and distort specific analyses, and, as a consequence, the respective outliers need to be removed.
Hence, the first preprocessing step in our single-cell workflow is a filtering procedure that identifies and removes barcodes with no or biased information, e.g., multiplets or damaged cells (cf. Figure 59A). To this end, we employ different quality control criteria that are based on the best practices guide by Luecken and Theis [320].
In general, measurements of multiple cells, such as doublets, can be identified by much higher counts or a larger fraction of expressed genes. Analogously, barcodes that represent damaged cells or empty droplets are often denoted by a small number of counts or a small fraction of expressed genes. Additionally, a high percentage of mitochondrial genes might also indicate damaged cells with broken membranes whose mRNA might have leaked out.
For all the described criteria, we provide filter procedures that can be applied to detect and remove outliers. The thresholds for the individual criteria depend on the experimental protocol and the analyzed cell types. For example, methods based on unique molecular identifiers (UMIs) often have smaller read counts than other approaches. Additionally, cells with a certain type, function, or state might naturally differ from other cells, e.g., cells involved in respiratory processes usually have a higher amount of mitochondrial genes [320].
For these reasons, it is mandatory to carefully select the thresholds for the different filter criteria. As default options, we employ permissive thresholds that are adapted to the different protocols. Additionally, we save the different quality metrics for each cell, such that users can assess them manually to identify further biases.
After the quality control steps, we assume that each barcode represents exactly one cell. Hence, both terms are used interchangeably in the following sections.

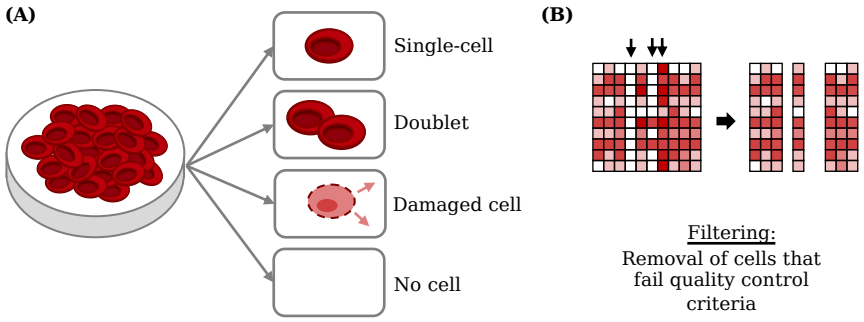


Figure 59: (A) Potential biases that can occur in a single-cell experiment: doublets, damaged cells, incomplete information, or no cell, e.g., due to empty droplets. (B) Overview of the filtering procedure. Cells that fail quality control checks are removed from the matrix.

5.4.1.2 *Normalization*

After quality control, the next preprocessing step in our single-cell analysis workflow is the normalization of the raw count values. This step is needed to ensure that expression values are comparable between and within samples. The choice of an appropriate normalization method depends on the sequencing protocol used. In general, the sequencing protocols can be categorized into two distinct strategies: methods that sequence the full length of a transcript and methods that only sequence the 3′ or 5′ end [518, 615]. In full length based approaches, multiple reads are usually generated from one transcript due to fragmentation. Generally, this means that longer transcripts produce more reads than shorter ones, and we need to apply normalization methods that account for this problem. For this purpose, we provide two approaches that adapt the raw counts by gene length and library size: TPM [554] (cf. Section 3.1.1.2) and GeTMM [495].
Since the length normalization step for both methods is optional, they can also be applied to normalize reads from protocols that only sequence the 3′ or 5′ end of transcripts and usually generate unique molecular identifiers (UMIs), i.e., only one read per transcript.
In all cases, we transform the normalized expression values by adding a pseudo-count and taking the logarithm ($\log_2$).

5.4.2 *Single-cell enrichment analysis*

In the next step, the normalized expression values can be used to conduct enrichment analyses for each cell (barcode). To this end, we select the most expressed genes for each cell, e.g., the 500 genes with the highest expression. For the selected genes, we calculate *over-representation analyses (*ORA*)* to detect active biological processes of each cell. In a subsequent step, the calculated enrichment results can be used to compare different groups of cells, such as disease and control. These groups can either be defined via annotations uploaded by the user or clusters calculated by our web service.
Since modern single-cell data sets regularly contain measurements for tens of thousands of cells, these steps can be very compute-intensive. Hence, we created an adapted ORA approach that utilizes an index structure containing p-values for predefined parameters instead of calculating the p-values from scratch. While this requires the creation of several specific indices, which is only possible for a small restricted parameter space, it reduces the total runtime drastically and facilitates the processing of huge data sets with the web service.
In general, this ORA approach is divided into two steps. In the first step, we precompute and save all p-values for the required test set and reference set sizes in a matrix. In the second step, this matrix is then used as a reference to obtain p-values in constant time.

5.4.2.1 *Calculation of p-value matrices*

Given a test set of size n, a reference set of size m, and a maximum category size l_{max}, we can calculate a matrix $P^{(l_{max}+1)\times(l_{max}+1)}$ that contains all p-values for categories of size $l \in [1, l_{max}]$ and $k \in [0, l_{max}]$ category members in the test set. Each entry $P[l, k]$ in this matrix represents the p-value for a category of size l and k test set members that belong to this category.
Instead of calculating all entries of this matrix independently, we employ a new dynamic programming approach. This is possible because the p-values for both *Hypergeometric test* and *Fisher's exact test* can be calculated recursively, i.e., $P(X \geqslant k) = P(X = k) + P(X \geqslant k+1)$.

PROOF The proof for this observation can directly be derived from the definition of $P(X \geqslant k)$:

$$P(X \geqslant k) = \sum_{i=k}^{min(n,l)} P(X = i) \tag{93}$$

$$\Leftrightarrow \quad P(X \geqslant k) = P(X = k) + \sum_{i=k+1}^{min(n,l)} P(X = i) \tag{94}$$

$$\Leftrightarrow \quad P(X \geqslant k) = P(X = k) + P(X \geqslant k+1) \ \square \tag{95}$$

Based on this observation, we can then use the following dynamic programming procedure to calculate a matrix containing all upper-tailed p-values efficiently. First, each row is initialized at position $P[l, d = min(n, l)]$ with the probability $P(X = d|m; n; l)$. The remaining entries in each row, are then calculated with the following recurrence formula:

$$P[l, k] = P[l, k + 1] + P(k|m; n; l), \tag{96}$$

where $k \in [d - 1, 0]$.

The pseudo-code for this algorithm is shown in Algorithm 1.
A dynamic programming approach for lower-tailed p-values can be derived accordingly.

Algorithm 1 Precomputation of upper-tailed p-values for an over-representation analysis with fixed reference and test set size.

```
1: Given the size of the reference set m, the size of the test set n, and the
   maximal category size l_max.
2: procedure PRECOMPUTE-P-VALUES(m, n, l_max)
3:     P[0..l_max, 0..l_max] = 1.0
4:     for l=0 to l_max do
5:         d = min(n, l)
6:         P[l, d] = P(d|m; n; l)
7:         for k=d-1 to 0 do
8:             P[l, k] = P[l, k + 1] + P(k|m; n; l)
9:     return P
```

5.4.2.2 *Index-based over-representation analysis (ORA)*

In a second step, the p-value matrix generated with the approach outlined above can now be utilized in an index-based ORA procedure. Here, the initial processing steps are equivalent to our original ORA approach (cf. Section 3.6.2), but instead of calculating a p-value for each test, we can extract the corresponding value from the reference matrix. The remaining processing steps stay the same.
In order to enable an efficient processing of hundreds of thousands of single-cells, we make sure that the same number of cells are used as a test set for each cell and the the same reference set is used for all over-representation analyses. Hence, we can use the same pre-computed p-value matrix for all computations. While these adjustments make the new ORA procedure less flexible than the standard approach, it improves the performance and makes the results of all cells better comparable. An overview of this method is depicted in Figure 60.

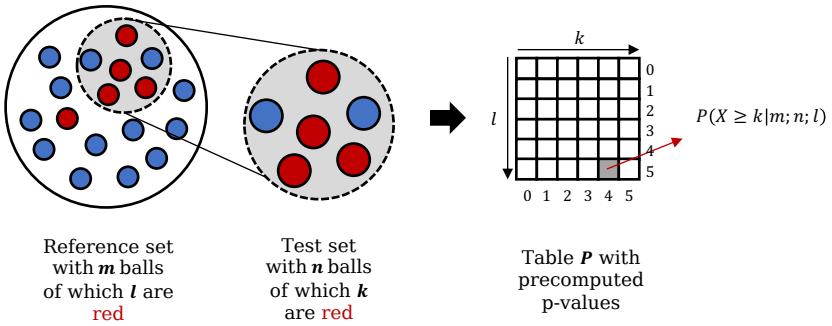


Figure 60: Overview of the index-based ORA approach. For a given reference set of size m and a test set of size n, the matrix P contains the p-values for all possible combinations of reference set and test set sizes. In particular, the entry $P[l, k]$ contains the pre-computed p-value for a category of size l and k test set members in this category.

5.4.2.3 *Enrichment-based group comparison*

In addition to the gene expression matrix, users can also upload a meta-data file containing annotations for each cell, such as cell type, tissue type, cluster assignments, or disease status. These annotations can be used to define groups and to assess if there are functional differences between the cells assigned to these groups.
For this purpose, we conduct a χ^2-test [440] that checks for each group and each biological category if the category is significantly more active in the investigated group than in all other cells. Each χ^2 test consists of two processing steps that are described in the following.
First, we have to define a 2×2 contingency table for each biological category. To this end, we divide the enrichment results for the processed category into four distinct classes that represent the entries of this table. Each entry represents the number of cells that belong to this class. Here, the rows divide the cells based on the group assignment, and the columns denote if the category was significantly enriched or not. An overview of this procedure is shown in Figure 61.

In the second step, we can then perform the χ^2-test to examine if group assignment and enrichment are independent or if the investigated group contains more cells with significant enrichment results than expected by chance:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

Here, n_{ij} indicates the number of cells and n^*_{ij} the expected frequency of this class based on the global distribution:

$$n^*_{ij} = \frac{(\sum_k n_{ik}) \cdot (\sum_l n_{lj})}{n},$$

Finally, a p-value for this test statistic can then be obtained from a χ^2-Distribution with 1 degree of freedom.

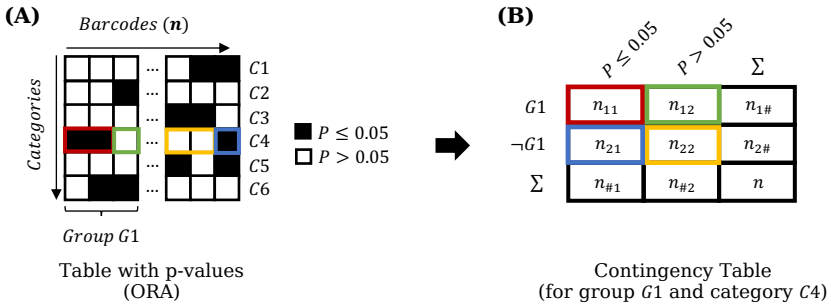


Figure 61: Creation of a contingency table for the χ^2-Test. (A) For each category in our single-cell enrichment results, we create one 2 by 2 contingency table. To this end, the cells are grouped based on two criteria, i.e., if a cell belongs to group G1 and if the considered category is significantly enriched. (B) The contingency table is then complemented by calculating the row and column sums.

5.4.3 *Group comparison and downstream analyses*

In addition to the new functionality for single-cells described in the previous sections, our framework offers different types of pseudo-bulk analyses for single-cell data. Here, instead of analyzing the measurements of each cell individually, cell groups with the same annotation are aggregated to so-called pseudo-bulks and then further processed. In the following, different analyses are described that follow this approach.

5.4.3.1 *Group comparison and enrichment analysis*

One of the most popular analyses in this context is the comparison of two cell groups of interest, e.g., disease vs. control. In order to compare these groups, our web service first aggregates the expression measurements in each group by calculating the mean expression values of each gene. Subsequently, our framework applies a fold-change to determine differences between the two considered groups. This analysis results in a score list that can directly be used in our standard enrichment analysis workflow (cf. Section 5.1).

In general, GeneTrail offers two different ways to access this functionality. Users can either upload the pseudo-bulk data directly to the standard analysis workflow or access the functionality via our single-cell toolbox.
Here, in order to conduct a group comparison, a user first has to select among the different cell annotations. Based on the selected annotation, the user can then define two cell groups that should be compared. After the groups have been selected, GeneTrail automatically conducts a group comparison and over-representation analyses for both the most over- and under-expressed genes. The results can then either directly be accessed on the single-cell page or via our detailed enrichment view (cf. Figure 53).

5.4.3.2 *Cell type prediction*

A special use-case of the workflow described in the previous section is that it can be applied for cell type prediction. For this purpose, a user has to select a cell annotation, e.g., a provided clustering. For each of the clusters, our web service then conducts a group comparison against the remaining cells. The resulting score list is then used in an enrichment analysis using the Wilcoxon rank-sum test (cf. Section 3.4.4.1 + 3.6). In this enrichment analysis, instead of biological processes, we utilize marker genes for known cell types, e.g., from The Human Protein Atlas [415]. Based on the most enriched categories, we can then estimate the underlying predominant cell type of each cluster. An overview of this approach is depicted in Figure 62.

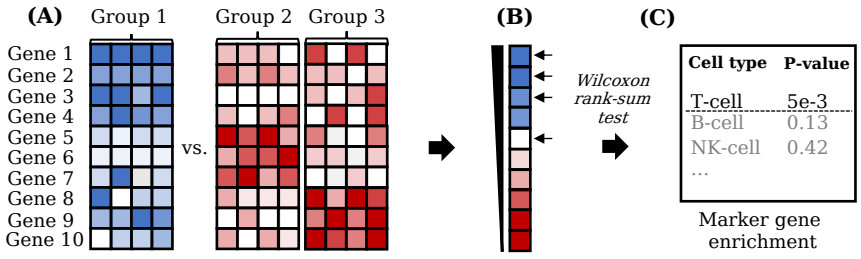


Figure 62: Overview of our cell type prediction approach. (A) Based on a given annotation, each group of cells is compared against all other groups. (B) We then apply a Wilcoxon rank-sum test to check if marker genes of a particular cell type are significantly enriched in the resulting list. (C) The enrichment results are then used to rank all cell types and to select the most likely one.

5.5 TIME-SERIES WORKFLOW

Author contributions

The time-series workflow was mainly designed and developed by Hans-Peter Lenhof and me. Anne Müller assisted with the implementation of the workflow. The complete list of contributors can be found in the author list of the respective publications [176, 177].

With the third version of GeneTrail, we also introduced the possibility to analyze time-resolved gene expression measurements. Here, we especially focus on the identification of biological processes whose activity changes over time. For this purpose, we developed a new pipeline that analyzes and compares the time courses of all features. First, a feature selection is conducted to remove features with no expression change in the analyzed time frame. For the remaining features, a clustering approach is applied that groups features with very similar expression curves. In the last step, an enrichment analysis is performed for each cluster to find the biological categories that are affected by the corresponding expression pattern.

In the following paragraphs, we first discuss the new methodology created for this pipeline. In Section 5.6.2, we then apply our tool to time-resolved transcriptomic and miRnomic profiles of CD4+ T cells to study processes with altered activity after T cell activation.

5.5.1 *Step 1 - Feature selection*

The goal of the time-series workflow is the identification of biological processes or signaling pathways whose activity changes during the analyzed time frame. To this end, the first step in our pipeline is the identification of interesting features. For this purpose, we employ several measures that assess the amount of fluctuation in a time interval. Only features that exceed a user-defined threshold for these measures are considered for subsequent analysis steps.

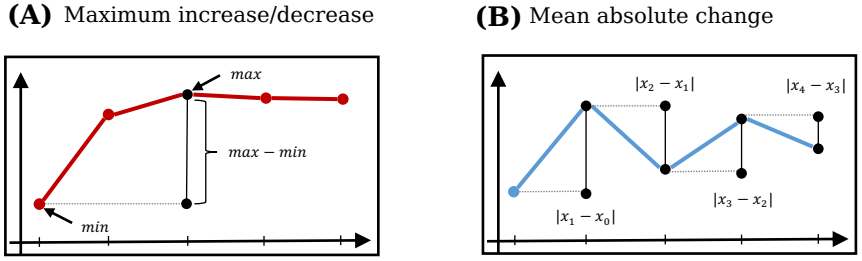


Figure 63: Feature selection measures for time-resolved expression data.

Given a series of expression measurements for a specific feature $x = \{x_1, ..., x_n\}$. The first measure calculates the maximum overall increase (or decrease) in expression.

$$d(x) = max(x) - min(x) \tag{97}$$

Similarly, we can define the maximum deviation from the first time point.

$$d(x) = max \begin{cases} |max(x) - x_1|, \\ |min(x) - x_1| \end{cases} \tag{98}$$

Both measures are well suited for cases, where we expect large changes in one direction.

In contrast to this, interesting curves could also be characterized by a high fluctuation. Hence, we can define a third measure as:

$$d(x) = \frac{1}{n-1} \sum_{i=1}^{n-1} |x_{i+1} - x_i| \tag{99}$$

Examples for both approaches are illustrated in Figure 63

5.5.2 *Step 2 - Clustering of time-resolved expression measurements*

The next step in our workflow is the identification of gene groups with very similar expression curves in the analyzed time frame. To this end, we use a two-stage clustering approach. In the first stage, we conduct a very strict clustering that creates many small groups with very similar expression patterns. In the second stage, the small clusters are further combined into larger groups, called "super-clusters". This approach structures the data into a hierarchy, where the super-clusters provide a general overview of expression trends in the data and the smaller clusters enable users to get a more in-depth view. An overview of the approach is shown in Figure 64.

5.5.2.1 *Distance and similarity measures*

One essential part of every clustering algorithm is the selection of a suitable distance or similarity measure. For time-resolved expression profiles, these measures need to consider the curve characteristics in the analyzed time frame and not only the distance of individual data points. In fact, for some analyses, we might want to group genes with a parallel time course even if they have a different expression level. Hence, in addition to standard measures such as correlation coefficients or the Euclidean distance (cf. Section 3.5 or Figure 65A), we implemented specialized measures for this purpose that are described in the following sections.

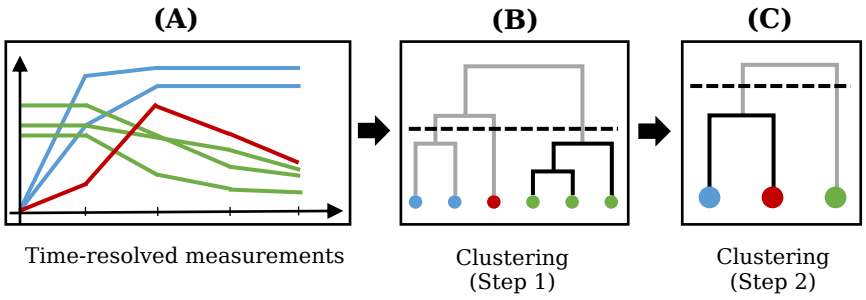


Figure 64: Overview of the two-stage clustering approach in the GeneTrail time-series workflow. In the first stage (A → B), a very strict clustering is applied to identify small and very similar clusters. In the second step (B → C), the clusters are further combined into super-clusters.

5.5.2.1.1 Adaptations of the Euclidean distance

According to Section 3.5, the Euclidean distance for two series of time-resolved measurements $p = (p_1, ..., p_n)$ and $q = (q_1, ..., q_n)$ is defined as:

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{100}$$

Since the standard definition uses the expression differences in each time point, it is not well suited to find curves with similar shapes that start on different expression levels (cf. Figure 65A). To overcome this, we propose the following two adaptations.

Shifted Euclidean distance

The first measure shifts the curves, such that the distance between them is minimized (cf. Figure 65B):

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - s - p_i)^2}, \tag{101}$$

where the optimal value for s is defined as:

$$s = \frac{1}{n} \sum_{i=1}^{n} (q_i - p_i) \tag{102}$$

The proof that s gives us the optimal value can be found in Appendix D.1.

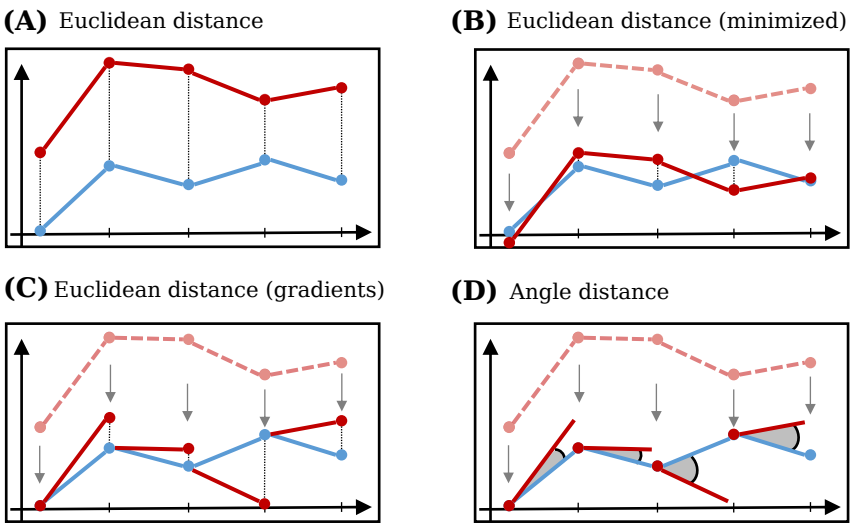


Figure 65: Overview of distance measures used to compare time courses: (A) Standard Euclidean distance, (B) Shifted Euclidean distance, (C) Euclidean distance between gradients, (D) Angle distance.

Euclidean distance between gradients

Alternatively, we can define the distance between p and q based on the gradient differences of all consecutive time points instead of the times directly (cf. Figure 65C):

$$d(p,q) = \sqrt{\sum_{i=1}^{n-1} ((q_{i+1} - q_i) - (p_{i+1} - p_i))^2} \tag{103}$$

5.5.2.1.2 Angle distance

The distance between two series of time-resolved measurements $p = (p_1, ..., p_n)$ and $q = (q_1, ..., q_n)$ can also be defined based on the angle differences of two consecutive points in the time series (cf. Figure 65D):

$$d(p,q) = \sum_{i=1}^{n-1} \theta((1, p_{i+1} - p_i), (1, q_{i+1} - q_i)), \tag{104}$$

Here the two points $(1, p_{i+1} - p_i)$ and $(1, q_{i+1} - q_i)$ define two vectors $\vec{u}$ and $\vec{u}$ that start in the origin $(0,0)$, and θ calculates the angle between these two vectors:

$$\theta(\vec{u}, \vec{v}) = \arccos\left(\frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \cdot \|\vec{v}\|}\right), \tag{105}$$

5.5.2.2 *Clustering*

Based on the statistical measures described in the previous section, we can then perform our two-stage clustering to identify gene groups with similar time courses. In the first stage, we apply a very strict clustering to identify small gene groups with very similar curve shapes. Since this usually results in many clusters that are hard to analyze manually, we conduct a second clustering that groups the initial clusters into larger classes, called "super-clusters". To this end, we first calculate the average (mean) time course of each initial cluster. For these aggregated time courses, we then conduct a second less stringent clustering to find the super-clusters. In practice, the super-clusters provide a general overview of the expression trends in the data. These can then be used as entry points for a more in-depth examination of the small clusters (cf. Figure 66).

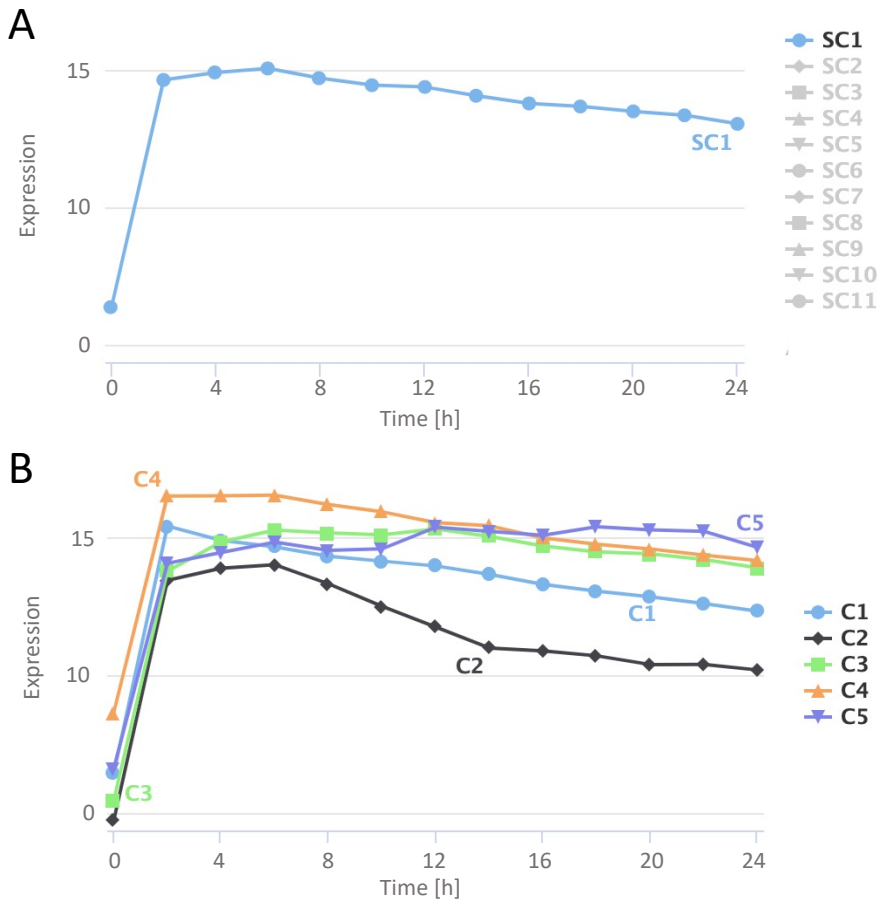


Figure 66: Screenshots of the GeneTrail time-series result visualization for the T cell activation data set (cf. Section 5.6.2). Depicted are the time points after T cell activation stimulus. (A) Average expression time course of super cluster SC1. (B) Average expression time courses of all clusters that belong to SC1.

In the current version of GeneTrail, both clustering stages can be conducted using any of the hierarchical clustering approaches described in Section B.1. However, instead of hierarchical clustering, alternative methods could also be applied, such as our clique partitioning approach described in the following paragraph [121].

Clique partitioning ILP

Author contributions

The ILP described in this paragraph was mainly developed by Kerstin Lenhof, Hans-Peter Lenhof, and me. It was applied in our publication "Quantitative and time-resolved miRNA pattern of early human T cell activation" [121] to study miRNA expression patterns during the T cell activation process. The complete list of contributors for this study can be found in the author list of the respective publication [121].

Besides standard hierarchical clustering, we additionally developed a graph-based approach that searches for an optimal clustering with a high similarity amongst all members [121]. This algorithm is an adaptation of the cutting plane clustering method by Grötschel and Wakabayashi [192].

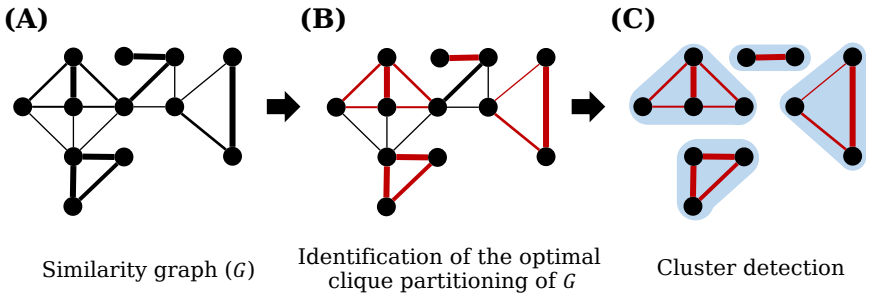


Figure 67: Overview of the clique partitioning approach. (A) The input of the clustering algorithm is a graph G, where each node represents a feature and each edge is weighted by similarity of the two connected features. (B) Our ILP then identifies the optimal set of edges that need to be removed from G to obtain the best clique partitioning. (C) The cliques then represent the final clusters.

The input of our algorithm is a similarity matrix S and a pre-defined threshold δ_S. Both S and δ_S are used to build an undirected graph $G = (V, E)$, where each vertex $v_i \in V$ represents a feature and each edge $(e_{ij} \in E)$ represents an entry $s_{ij} \in S$ iff $s_{ij} > \delta_S$. The goal of our method is to find the best possible clique partitioning of G, i.e., the set of fully connected subgraphs that have the highest total edge weight (similarity). This is achieved by removing the edge set with the smallest possible weight and simultaneously producing a valid

clique partitioning of G. An overview of this approach is shown in Figure 67.
We formulate this clustering approach as an integer linear program (ILP). For this purpose, we first define a binary decision variable y_{ij} that specifies if an edge belongs to the edge set that is removed from the graph (E_{cut}).

$$y_{ij} = \begin{cases} 1, \text{ if } e_{ij} \in E_{cut} \\ 0 \end{cases} \tag{106}$$

Based on this, we can define an objective function that minimizes the weight of E_{cut}.

$$\min \sum_{e_{ij} \in E} s_{ij} y_{ij} \tag{107}$$

Next, we need to make sure that the selected solution is a valid clique partitioning of the graph, i.e., that each connected component is a fully connected subgraph. To this end, we first define an auxiliary variable z_{ij} that specifies edges that remain part of the graph.

$$z_{ij} = \begin{cases} 1 - y_{ij}, \text{ if } e_{ij} \in E \\ 0 \end{cases} \tag{108}$$

For all edges with $z_{ij} = 1$, we now need to ensure that they are part of a clique. We can achieve this by verifying the clique property for all triplets of vertices. To this end, we need to consider four different cases (cf. Figure 68). However, we only need to add additional constraints for the first two (A + B). The two remaining cases cannot break the clique property and, hence, can be neglected.
In the first case, all three vertices are connected and form a cycle. In the following, we denote the set of all cycles with three vertices as C_3 (cf. Figure 68A). For each cycle in C_3, we have to make sure that all three edges remain part of the graph or at most one. This can be guaranteed by the following constraints:

$$z_{ij} + z_{jk} - z_{ki} \leqslant 1 \; \forall i, j, k \in C_3 \tag{109}$$
$$z_{ij} + z_{ki} - z_{jk} \leqslant 1 \; \forall i, j, k \in C_3 \tag{110}$$
$$z_{jk} + z_{ki} - z_{ij} \leqslant 1 \; \forall i, j, k \in C_3 \tag{111}$$

Additionally, we need to add a similar constraint for all triplets that represent a path of length two, which we denote P_2 (cf. Figure 68B). In this case, at least one has to be removed, since they cannot be part of the same clique.

$$z_{ij} + z_{jk} \leqslant 1 \; \forall i, j, k \in P_2 \tag{112}$$

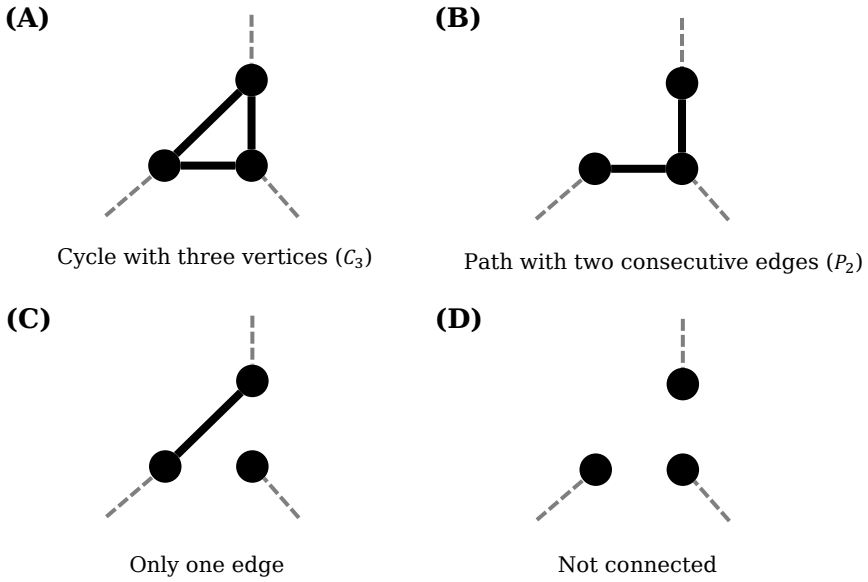


Figure 68: Examples of the four cases that need to be considered to ensure the clique property in our approach. However, since (C) and (D) cannot break this property, they can be neglected.

Our ILP formulation differs from the original one by Grötschel and Wakabayashi [192] in the way the clique partitioning is ensured for edges that are part of the final solution. In both approaches, all possible triplets of vertices are considered to ensure the clique property. However, in our case, constraints are only added to the ILP if a vertex triplet forms a cycle (C_3) or a path (P_2). In contrast, the original ILP uses the three cycle constraints for all possible triplets of vertices. While, in theory, this achieves the same goal, without the need to distinguish the different cases depicted in Figure 68, it results in much more inequalities that need to be considered to solve the optimization problem than the explicit definition. Hence, one advantage of our adapted formulation is that this often reduces the required computation time. Moreover, our approach makes it much easier to run the algorithm iteratively, e.g., by adding additional edges to the graph and resolving the ILP based on the previous solution.

5.6 RESULTS

In this section, we present the results of two studies we conducted to highlight the capabilities of GeneTrail. Thereby, we focus on the single-cell and time-series functionality of our web service. Additionally, the standard functionality is demonstrated in our Wilms' tumor study (cf. Chapter 9).

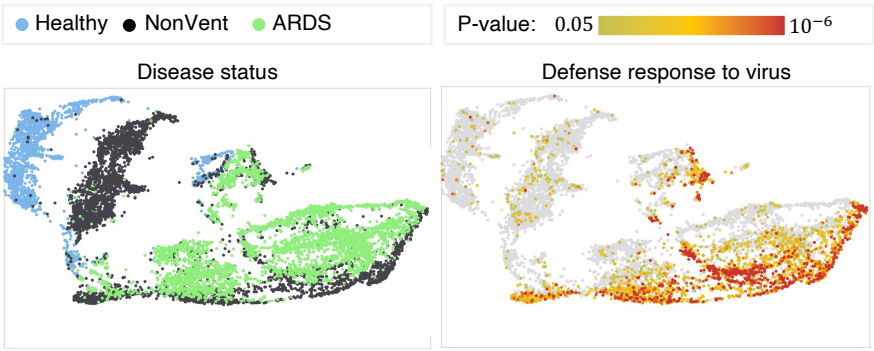
5.6.1 *Single cell analysis of monocytes from peripheral blood of* COVID − 19 *patients*

Author contributions

The analysis described in this section is based on our publication "GeneTrail: A Framework for the Analysis of High-Throughput Profiles" [176]. The data analysis was mainly conducted by Nico Gerstner, Hans-Peter Lenhof, and me. The complete list of contributors can be found in the author list of the respective publications [176].

Since the beginning of the year 2020, the coronavirus disease 2019 (COVID-19) has become a global pandemic. This highly contagious respiratory illness is caused by the coronavirus SARS-CoV-2 [399, 598]. According to the World Health Organization, nearly 250 million people were infected by October 2021 [578]. The fast spread of this disease was possible since asymptotic or presymptomatic cases are postulated to have been the driver of the epidemic [169, 362]. A study in the Chinese population showed that around 80% of people only have mild or moderate symptoms, while the remaining 20% seem to have severe or critical courses of the disease [589]. Many of the severe critical cases exhibit a highly deregulated activity of the immune system [178], which amongst others includes highly increased concentrations of pro-inflammatory cytokines [145]. In current literature, several putative causes for the heavily altered immune activity are actively discussed, such as pathogenic T cells [614], inflammatory monocytes [195], or myeloid-derived suppressor cells [10].
In this section, we demonstrate the capabilities of GeneTrail's single-cell workflow by analyzing a single-cell RNA-seq data set of CD14 monocytes from peripheral blood of COVID-19 patients and healthy controls [581]. This data set contains gene expression profiles of 10,339 cells from blood samples of seven hospitalized patients with COVID-19 and six healthy controls. Of the seven hospitalized patients, four were diagnosed with acute respiratory distress syndrome (ARDS) and required mechanical ventilation.
Here, we use GeneTrail to find biological processes that distinguish COVID-19 patients with ARDS (ARDS group) from patients that required no mechanical ventilation (NoVent group) and healthy controls (Healthy group). To this end, we calculated ORAs for the 500

most expressed genes in each cell and then compared the three groups using a χ^2-Test as described in Section 5.4. A description of all processing steps and the complete list of parameters can be found in Appendix E.3. In the subsequent paragraphs, we discuss some of the most striking differences between the three analyzed groups: ARDS, NonVent, and Healthy. This may help to elucidate putative causes or contributing factors of the more severe courses of disease in the ARDS group.



Category	Healthy	NonVent	ARDS
defense response to virus	5.44e-131	2.56e-129	1.01e-5
regulation of viral genome replication	2.04e-122	3.69e-105	0.0081
negative regulation of viral genome replication	6.58e-47	4.29e-74	4.92e-11
negative regulation of viral process	4.49e-60	1.33e-72	1.90e-6
response to virus	1.87e-246	1.29e-64	1.47e-25

Figure 69: Screenshot of the GeneTrail single-cell result page. On the bottom, the most significantly enriched biological processes in the NonVent group of our COVID-19 analysis are depicted. The table depicts the results the χ^2-tests used for group comparison and the background color indicates if the corresponding p-value is significantly enriched (red) or depleted (green). On the top, UMAP representations of the data set are plotted. The figure on the left is colored according to the group assignment, and the figure on the right is colored with respect to the p-value of the category "defense response to virus". This figure is adapted from [176].

Amongst the top results in our analyses, we obtain the process "defense response to virus" as well as further related categories. As depicted in Figure 69, these processes seem to be highly active in cells of the NonVent group, less active in cells of the ARDS group, and inactive in Healthy cells. In general, we would expect that the immune response to a virus infection is much more active in COVID-19 patients compared to healthy controls. However, the decreased activity of these processes in the ARDS could be an indicator that the adaptive immune response might be impeded in these patients. Similar observations have been frequently reported in severe cases [23,

243]. In this context, it has also been discussed that an impaired antiviral response causes a higher virus load in the blood of respective patients [198]. Hence, building upon these results, we focus in the following discussions on other critical immune-related processes that might help us to further elucidate the status of the adaptive and the innate immune system in the different groups.

First, we analyzed biological categories associated with interferon signaling or the response to interferons (IFNs) of type I, II, and III. Here, especially IFNs of type I and corresponding interferon-stimulated genes (ISGs) have important roles in the antiviral response. Analogously to our previous observations, the respective processes are significantly more active in the NonVent group. This phenomenon has also been repeatedly observed in severe cases of COVID-19 [1, 198, 292]. Here, a deficiency of type I IFNs in severe cases is also often observed in concordance with elevated activities of the TNF and NFκB signaling pathways. Corresponding categories show the expected behavior in our enrichment result list.

We also detected several other processes that confirm our observation about the reduced adaptive immunity in patients of the ARDS group. This includes biological categories involved in antigen presentation via MHC class II, which are also significantly depleted in cells of patients with ARDS. Recent studies also discussed this as a potential indicator of disease severity [303, 502]. Accordingly, we also observe a reduced expression of several MHC class II member genes, in particular, HLA-DRA, which has been reported as a marker for respiratory failure in severe cases of COVID-19 [178].

With respect to the innate immune system, we identified an elevated macrophage activity in the ARDS group. This is further confirmed by an enriched endocytosis and phagocytosis function in cells of patients with ARDS. An increased macrophage activity in COVID-19 patients also seems to be linked to a higher mortality [38]. Our result also indicated that cells from patients with ARDS might show a significant enrichment of processes related to motility, migration, and chemotaxis. This could potentially also be associated with the increased TNF activity in those cells [361].

In summary, our results suggest that the innate immune system is much more active in cells of patients with ARDS compared to cells of patients that did not require mechanical ventilation and healthy controls, while the adaptive immune system might potentially be impaired. Similar conclusions have been drawn by several studies that compared mild and severe cases of COVID-19 [23, 243].

5.6.2 *Time series analysis of the T cell activation process*

Author contributions

The data set analyzed in this section was published as part of our publication "Quantitative and time-resolved miRNA pattern of early human T cell activation" [121]. The analysis described in this section is based on our publication "GeneTrail 3: advanced high-throughput enrichment analysis" [177]. It was mainly conducted by Hans-Peter Lenhof and me. The complete list of contributors can be found in the author list of the respective publications [177].

T-helper cells or CD4+ cells are a group of lymphocytes that control the adaptive immune response. CD4+ can recognize specific antigens via the MHC class II receptors on the surface of antigen-presenting cells (APCs). After activation, CD4+ cells secrete different types of cytokines to coordinate the response of other immune cells, such as cytotoxic T cells or macrophages.

In this section, we demonstrate the capabilities of our time-series workflow by analyzing a data set of time-resolved gene expression profiles of CD4+ cells that were in vitro activated [121]. In total, the data set consists of 39 microarrays that measure the gene expression of 13 different time points in a 24 hour period after the T cell activation signal. For each time point, three replicates were created. The initial time points (T0) represent the expression before the cells were in vitro activated. Subsequently, the expression was measured after two-hour intervals, from 2 to 24 hours, after the initial activation signal. Here, we use the GeneTrail time-series functionality to analyze biological processes and signaling pathways that are induced by the T cell activation process. To this end, we first aggregated the replicates for each time point using the median of the three microarrays. Subsequently, we applied a hierarchical clustering with complete linkage and Euclidean distance between the gradients of consecutive time points to the genes with the overall highest increase in expression (cf. Section 5.5). All processing steps and the complete list of parameters can be found in Appendix E.4.

In total, we obtained 422 clusters that are grouped in 21 super-clusters. From the 21 super-clusters, we selected three that have different response to the T cell activation stimulus and, hence, have different expression time curves, i.e, SC1, SC7, and SC16 (cf. Figure 70A).

In the following paragraphs, we first discuss the biological processes and corresponding genes that are associated with each super-cluster individually, and then we use the associated expression time curves to examine the chronological order in which biological processes seem to be executed.

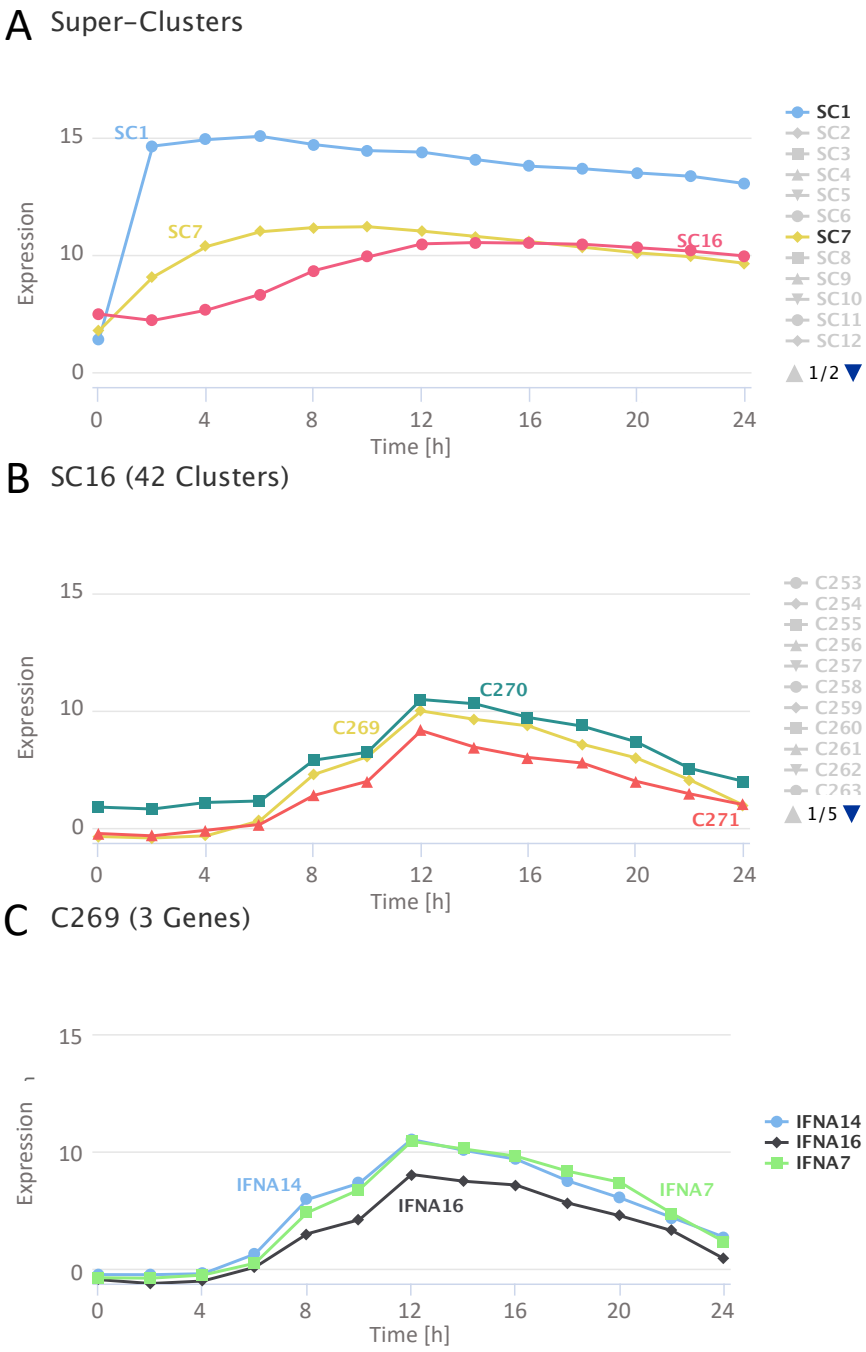


Figure 70: Screenshots of the GeneTrail time-series result visualization for the T cell activation data set. Depicted are the time points after T cell activation stimulus. (A) Three of the 21 super-clusters detected by GeneTrail. (B) Three of the 42 clusters that belong to SC16. (C) Member genes of cluster C269.

Super-cluster SC1 contains the genes with the highest overall increase in expression after two hours: CSF2, EGR2, IFNG, IL2, IL3, and TNF. After the initial increase, the overall expression level is slightly decreasing, although it stays constantly high. Most of these genes are known products of the "T cell receptor signaling pathway" (cf. KEGG [394], hsa04660) and directly involved in T cell activation. Moreover, we found various enriched categories that are directly associated with early immune response processes, including T cell costimulation or T cell differentiation. Additionally, we observe an enrichment of T cell activation directly, including several known hallmarks of this process, e.g., cytokine signaling, proliferation, and metabolic processes [89]. Among the top results, we also find categories that indicate a phosphorylation of STAT proteins and the positive regulation of the JAK/-STAT pathway.

The second super-cluster SC7 contains genes with a direct increase in expression that continues until 8 hours after the activation stimulus (cf. Figure 70A). It contains many cytokines (e.g., IL10, IL22, or LIF), transcriptional regulators with important roles in T cells (e.g., IRF4, TBX21, or VDR), and several members of the solute carrier family of vesicular transporter (e.g., SLC32A1, SLC3A2, or SLC7A1). Accordingly, amongst the most significant processes in our enrichment results, we found several processes related to the cytokine-mediated signaling pathway, leukocyte migration, chemotaxis, regulation of cytokine production, and transmembrane transporter activity.

In contrast to super-clusters SC1 and SC7, which seem to feature biological processes with an immediate increase in expression, SC16 shows a delayed response that starts after 4 hours and has a peak after 12 hours. The cluster contains many members of the interferon (IFN) family, i.e., 21 type I IFNs and 8 type II IFNs (cf. Figure 70C). Consequently, we observe an enrichment of the interferon signaling pathways as well as many processes involved in the activation of other leukocytes (e.g., B cells, T cells, or NK cells), or the defense response to external stimuli (e.g., dsRNA, virus, organic compounds, etc.).

Based on the enrichment results for each cluster discussed above and the corresponding expression time courses, we can now also infer putative chronological orders in which biological processes are executed. It seems that in general, the initial activation of the T cell receptor causes a highly elevated production of several key cytokines (e.g., IFNG, IL10, or TNF) and regulators of the early immune response (e.g., IRF4 or TBX21) that immediately stimulate key signaling pathways and the production of further cytokines and vesicle transporter which can help to secrete them. For example, it is established that ele-

vated levels of IFNG might stimulate interferon type II production (cf. WikiPathway [256], WP619) and that IL10 induces Jak-STAT signaling pathways which then promotes interferon type I production (cf. WikiPathway [256]). Finally, the production of the different interferon types seems to initiate an immune response to the provided activation stimulus.

In conclusion, the analysis described above demonstrates that our web service is well equipped for the identification of genes with similar expression time courses, the assessment of corresponding biological processes, and their chronological order.

5.7 SUMMARY, DISCUSSION, AND CONCLUSION

In this chapter, we presented the GeneTrail web service, a powerful toolbox for enrichment and network analysis. Compared to other tools (cf. Figure 47), our framework stands out by providing a rich functionality with highly efficient C++ implementations and interactive visualizations that can be used to analyze our comprehensive collections of biological categories for 15 organisms (cf. Appendix G). Moreover, our web service is the only one of the discussed tools that can directly be applied to analyze both time series and single-cell experiments.
Furthermore, GeneTrail is closely integrated with its sister projects DrugTargetInspector [474] and ClinOmicsTrail [475], which provide rich functionality for the analysis of molecular profiles of tumors. They were designed to support clinicians in the selection of personalized therapies for cancer patients (cf. Appendix C).
Nevertheless, there are still elements that can be improved or extended. For example, we plan to extend the support for additional omics types, e.g., glycomic, lipidomic, or metabolomic measurements. Accordingly, the functionality for the integrative analysis of multiple data types could also be extended. Here, in particular, single-cell multimodal omics data, where different data types are measured in the same cells, seem to have a high potential to gain novel insights into cell biology [527].
However, the rich functionality that can be applied in various application scenarios already makes GeneTrail one of the most comprehensive tool suites for the analysis of molecular high-throughput profiles and set it apart from other approaches.

6

MIRPATHDB - THE MIRNA PATHWAY DICTIONARY

Author contributions

This chapter describes the miRNA and pathway dictionary miRPathDB and is based on the respective publications [31, 33, 529, 530]. The original dictionary of miRNAs and their putative target pathways was created by Christina Backes, Andreas Keller, and Hans-Peter Lenhof in 2010 [31]. miRPathDB is an updated web-based version of this database. It was mainly developed by Christina Backes, Andreas Keller, Fabian Kern, Hans-Peter Lenhof and me. The data processing was conducted by Christina Backes, Fabian Kern and me. The web application was mainly implemented by me with further contributions by Daniel Stöckel and Lara Schneider. The complete list of contributors can be found in the author list of the respective publications [31, 33, 529, 530].

The analysis of signaling pathways is a crucial task in biomedical research that can help to advance our understanding of cellular mechanisms or even pathological processes. Key elements in the control of many biological processes are small regulatory RNA molecules, called miRNAs (cf. Section 2.1.4.3). These non-coding RNAs have been shown to orchestrate many important cellular functions by inhibiting the expression of their target genes in many different organisms [40, 246] (cf. Section 2.1.5.4). Due to their potential to control nearly all biological processes they are intensively studied and the total number of discovered miRNAs and available target genes is constantly increasing [15, 253].

One crucial task in current miRNA research is the functional analysis of these regulators. In particular, two important questions are often studied: (1) which set of miRNAs is involved in the regulation of a particular biological process (pathway-centric view) or (2) which biological processes are controlled by one specific miRNA (miRNA-centric view) [530]. To answer both questions, a variety of tools and databases have been proposed. Amongst the miRNA-centric resources are the databases miRTar [222] and miRSystem [317] that link miRNAs to signaling pathways, the miRNet web service that provides a network-based approach to study the function of miRNAs [143], and miTALOS, a tool for tissue-specific regulation of biological pathways [418].

Amongst the primarily pathway-centric resources are enrichment-based tools like the miRNApath R-package [96] and BUFET [600] that use the target genes of a miRNA set to identify associated cellular processes, and PolymiRTS, which assigns SNPs in target genes to associated pathways of disease phenotypes [57]. There are also several resources that can be used to investigate both questions, like miRNApath database [91], the CORNA R-package [588], and the DIANA-miRPath web service [551].
In this chapter, we present miRPathDB [33, 529, 530] a web-based dictionary on miRNAs, target genes, and target pathways for human and mouse. In addition to experimentally validated miRNAs that are supported by all tools described above, our database also contains information about putative miRNA candidates from the miRCarta database [28]. miRPathDB not only compiles the data from a miRNA- as well as pathway-centric view, but also provides several interactive analysis tools that can be used to evaluate the database content in a user-specific context. In the following sections, we describe the resources and methodology used to create miRPathDB (cf. Section 6.1), the database content (cf. Section 6.2), and analysis tools (cf. Section 6.3).

6.1 MATERIALS AND METHODS

miRPathDB depends on several third-party resources that were processed to create our database. These resources and the different processing steps are described in the following paragraphs.

6.1.1 *miRNA and miRNA candidates*

Our database uses genomic positions and sequence information in form of primary transcripts, mature sequences, and seed sequences for all human and mouse miRNAs from miRBase (V22) [281] and all human miRNAs and putative miRNA candidates from miRCarta [28] (V1.1). In the following, we use the term miRNA to refer to both validated miRNAs and miRNA candidates.

6.1.2 *Target genes and miRNA-target interactions (MTIs)*

In order to obtain the target genes of each miRNA, we used both experimentally validated target genes from miRTarBase [228] (Version 7) and predicted target genes. For the former, we created two target gene sets for each miRNA: (1) target genes with strong experimental evidence and (2) target genes with any experimental evidence. To predict the target genes for all miRNAs and miRNA candidates, we used TargetScan (Version 7.1) [8] and MiRanda (Version 3.3a) [138]. We applied both algorithms with default parameters to scan

the 3′ untranslated regions (UTRs) of all target genes using the seed sequences of all miRNAs. As UTRs, we use the curated annotations from *targetscan.org*. We then combined the predictions of both algorithms to create two consensus sets: (1) the intersection and (2) the union of all predictions. This is a common strategy to balance sensitivity and specificity of the predictions [56].
In the end, we obtained up to four target gene sets for each miRNA that were further processed:

- Validated targets with strong experimental evidence
- Validated targets with any experimental evidence
- The intersection of all target predictions
- The union of all target predictions

6.1.3 *Target pathways*

The target gene sets for each miRNA were used to identify putative target pathways. To this end, we employed the enrichment analysis functionality of GeneTrail (cf. Section 5). For each biological category from Gene Ontology [100], KEGG [394], Reactome [142], and WikiPathways [256], we conducted an over-representation analysis to check if there is a significant overlap with the target genes of a particular miRNA. As a reference set, we used all genes for which miRNA target information was available. All resulting p-values were adjusted using the method proposed by Benjamini and Hochberg [44] (cf. Section 3.3.2). Finally, we assumed that a biological category is controlled by a specific miRNA if the adjusted p-value of the corresponding ORA is below 0.05.

6.1.4 *Comparison between miRNAs*

In general, the miRNA sequences are highly conserved [390], and miRNAs derived from a common ancestor can even be grouped into larger miRNA families that often have similar sequences and similar molecular functions [616]. Accordingly, miRNA with similar (seed-)sequences often share similar target genes and target pathways [529]. Consequently, this information can also help to study putative functions of miRNA candidates.
Hence, in addition to sequence, target gene, and target pathway information, miRPathDB also provides similarity information for all miRNA pairs. To this end, we analyzed and compared the mature sequence, seed sequence, target gene sets, target categories, and genomic positions of all miRNA pairs. In the following, we describe the measures we used for this purpose.

6.1.4.1 *Sequence similarity*

For both, seed and mature sequence of all miRNA pairs, we use a similarity measure that is based on the hamming distance. Given the alphabet $\Sigma = \{A, C, G, T\}$ and two sequences $s, t \in \Sigma^n$, it can be calculated as the number of positions with different letters [200, 579].

$$D_{Hamming}(s, t) = |\{i \in \{1, ..., n\} | s_i \neq t_i\}| \tag{113}$$

Based on this distance, we then defined the following similarity:

$$Similarity(s, t) = 1 - (\frac{D_{Hamming}}{n}) \tag{114}$$

6.1.4.2 *Similarity of target genes and target pathways*

For the different sets of target genes and target pathways, we use the Jaccard coefficient (JCC) [240]. For two sets X and Y, it can be defined as:

$$JCC(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{115}$$

6.1.4.3 *Genomic distance*

For all miRNA pairs located on the same chromosome, we also calculated the genomic distance as the number of bases between their loci.

6.2 DATABASE CONTENT

We designed miRPathDB as a web resource for biomedical researchers that are interested in links between miRNA, their target genes, and target pathways. The data can be accessed in three different ways: (1) by downloading raw or processed data in different text-based formats, (2) by using our web-based API, or (3) by using our interactive web interface. The latter compiles the provided information either from a miRNA-centric or a pathway-centric point of view. Both viewpoints are described in the following sections.

6.2.1 *miRNA-centric view*

The miRNA-centric view of our database provides different levels of information for individual miRNAs. The page of a particular miRNA can either be accessed via the overview page or queried in one of the search bars. Most information is provided as responsive, searchable, and interactive tables that can be downloaded in different formats (Excel, PDF, CSV).

6.2.1.1 *General information and links to external resources*

On top of each page, general information about the miRNA is displayed. This includes information about the miRNA family, precursor, and sequence information (cf. Figure 71). Additionally, links to external resources are provided, such as miRBase [281], miRCarta [28], the miRNA tissue atlas [318], and miRTargetLink [261].

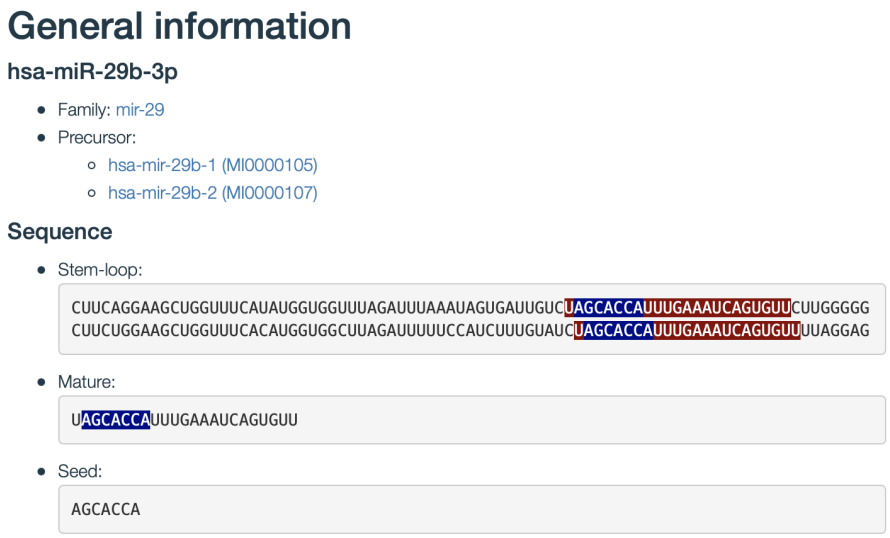


Figure 71: Screenshot of miRPathDB: Information about stem-loop, mature, and seed sequences of hsa-miR-29b-3p.

6.2.1.2 *Target gene sets*

Below the general information, miRPathDB contains a table that lists all target genes and the respective evidence, i.e., if the target was experimentally validated or predicted (cf. Figure 72).

Target	Evidence
A1BG	predicted (union)
A1CF	predicted (union)
A4GALT	predicted (union)
AADACL4	predicted (union)
AAK1	predicted (intersection) + predicted (union)

Figure 72: Screenshot of miRPathDB: Information about target genes of hsa-miR-29b-3p.

6.2.1.3 *Target pathways*

The main focus of our database is to provide putative links between miRNAs and associated biological processes or signaling pathways. As shown in Figure 73, this information is provided as an interactive table that provides custom filter procedures for each column. In the text box below each column, users can type text or simple equations to select subsets of the data or to search for specific target genes or pathways.

Database	Pathway	Evidence	Hits	Expected hits	P-value
Reactome	Extracellular matrix organization	experimental (any)	41	4.444	2.47e-25
KEGG	Focal adhesion	experimental (any)	32	3.294	1.40e-20
KEGG	PI3K-Akt signaling pathway	experimental (any)	38	5.158	1.56e-20
Reactome	Extracellular matrix organization	experimental (strong)	32	3.935	3.92e-20
Reactome	Assembly of collagen fibrils and other multimeric structures	experimental (any)	20	0.924	6.98e-20
		experimental	>=20		<0.0001

Figure 73: Screenshot of miRPathDB: Information about target pathways of hsa-miR-29b-3p. The results are filtered with respect to three filter criteria: (1) experimental evidence, (2) more then 20 target genes, and (3) a p-value smaller than 0.0001. On the miRPathDB website, this table also depicts the target genes. Here, due to space constraints, this columns was removed.

6.2.1.4 *Similarity to other miRNAs*

At the bottom of each page, miRPathDB contains an additional table that compares the miRNA to all other miRNAs and miRNA candidates in our database. Each column contains the information about a different comparison (cf. Figure 74). This can be used to study relationships between miRNA pairs and the different types of information, i.e., sequence similarity, genomic positions, similarities between target gene sets, or target pathways (cf. Section 6.1).

miRNA	Sequence similarity (seed)	Sequence similarity (mature)	Chromosomal distances	Jaccard coefficient (target genes - prediction intersection)	Jaccard coefficient (target genes - prediction union)
m-81	1.000	1.000	0.00e+0	1.000	1.000
m-82	1.000	1.000	0.00e+0	1.000	1.000
hsa-miR-29b-3p	1.000	1.000	0.00e+0	1.000	1.000
m-51	1.000	0.905	564.0	0.917	0.930
hsa-miR-29c-3p	1.000	0.864	564.0	0.928	0.936

Figure 74: Screenshot of miRPathDB: Similarities between hsa-miR-29b-3p and all other miRNAs.

6.2.2 *Pathway-centric view*

The data contained in our database can also be accessed from a pathway-centric point of view. Here, each biological category has its own web page. These pages consist of a single table that contains all miRNAs with a significant enrichment of the respective biological process or signaling pathway (cf. Figure 75).

Database	miRNA	Evidence	Hits	Expected hits	P-value
miRCarta	m-17786	predicted (intersection)	42	23.5922	0.002
miRBase	hsa-miR-122-3p	experimental (strong)	2	0.118732	0.004
miRBase	hsa-miR-369-5p	experimental (strong)	2	0.178097	0.010
miRBase	hsa-miR-558	experimental (strong)	2	0.178097	0.010
miRBase	hsa-miR-433-5p	experimental (strong)	2	0.178097	0.010
miRBase	hsa-miR-371a-3p	experimental (strong)	2	0.237463	0.019
miRBase	hsa-miR-29c-5p	experimental (strong)	2	0.237463	0.019
miRCarta	m-4802	predicted (intersection)	202	158.361	0.020
miRBase	hsa-miR-26a-1-3p	experimental (strong)	3	0.474926	0.020
miRBase	hsa-miR-3127-5p	experimental (strong)	2	0.237463	0.024

Figure 75: Screenshot of miRPathDB: Overview of all miRNAs that control the "PI3K-Akt signaling pathway" from the KEGG database.

6.3 ANALYSIS TOOLS AND EXAMPLE APPLICATIONS

Apart from the database content itself, we also created interactive analysis tools that are described in the following paragraphs.

6.3.1 *Custom pathway heatmaps*

As shown in the previous section, miRPathDB provides links between individual miRNAs and putative target pathways. This is valuable information to study which molecular functions are controlled by a particular miRNA or vice versa. Building on this, a logical next step is to consider more than one miRNA and to assess if they might be involved in common biological functions. For this purpose, we created an interactive pathway heatmap visualization that can be used to investigate this question.
The input for this tool is a set of user-defined miRNAs that should be compared. miRPathDB then queries our database to find all biological categories that are significantly enriched for the target genes of at least one of the input miRNAs. Next, the p-values of the respective categories are used to create a miRNA $\times$ category matrix, where each entry represents the $\log_{10}$-transformed p-values. Our database then conducts a hierarchical clustering with Ward's method (cf. Section B.1) and Euclidian distance (cf. Section 3.5) for both rows and columns of the matrix to group similar miRNAs and pathways respectively. The final matrix is then visualized in the browser as an interactive heatmap plot (cf. Figure 76).

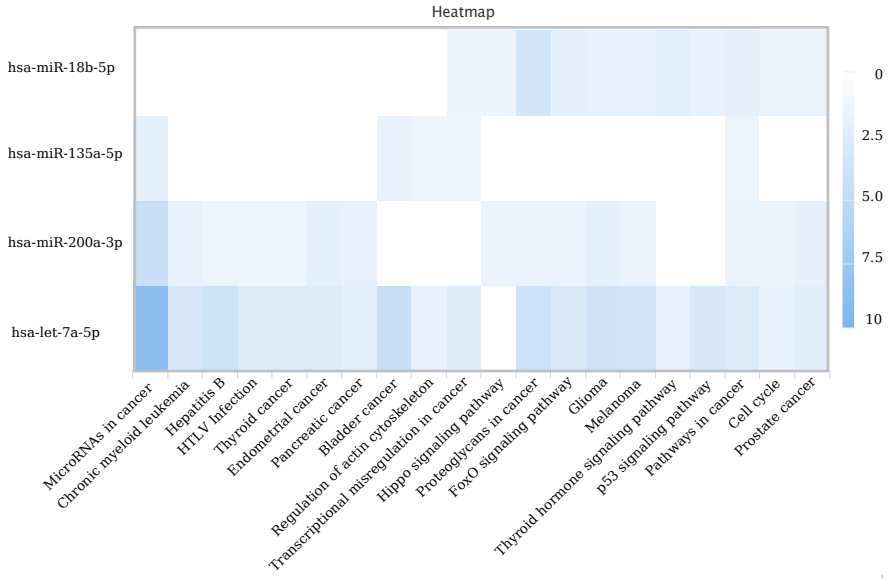


Figure 76: Screenshot of miRPathDB: Pathway heatmap of hsa-miR-18b-5p, hsa-miR-135a-5p, hsa-miR-200a-3p, and hsa-let-7a-3p and biological categories from KEGG with strong experimental evidence.

6.3.2 *Maximum targetome coverage analysis*

A further interesting task in functional miRNA research is the identification of a small number of miRNAs that are sufficient to completely regulate a given biological process or signaling pathway. To study this issue, we created a new tool that is based on the maximum coverage problem. In the following, we first define the theoretical problem and then discuss its application to miRNA-target interactions (MTIs).

6.3.2.1 *Maximum coverage problem*

Given a set of l miRNAs $M = \{m_1, ..., m_l\}$ and n target genes $G = \{g_1, ..., g_n\}$ of which a subset $T_i \subseteq G$ is regulated by a specific miRNA $m_i \in M$. Then the maximum coverage problem searches for a set of k miRNAs that target the maximal number of genes in G. It can be formulated as an integer linear problem (ILP). To this end, we first define two binary variables x_i and y_j.

x_i specifies for each miRNA if it is selected or not.

$$x_i = \begin{cases} 1 \text{ if miRNA } m_i \text{ is selected} \\ 0 \end{cases} \tag{116}$$

y_j indicates for any gene $g_j \in G$ if it is targeted by any of the k selected miRNAs.

$$y_j = \begin{cases} 1 \text{ if } g_j \text{ is targeted} \\ 0 \end{cases} \tag{117}$$

Based on these variables, we can define an objective function that maximizes the number of selected genes.

$$\max \sum_{g_j \in G} y_j \tag{118}$$

Apart from this, we also need two constraints, which ensure that a valid solution is selected. The first constraint makes sure that only solution with at most k miRNAs are selected.

$$\sum_{i=1}^{n} x_i \leqslant k \tag{119}$$

The second constraint ensures that each gene $g_j \in G$ is only selected ($y_j = 1$) iff it is targeted by at least one miRNA.

$$\sum_{\forall m_i \in M : g_j \in T_i} x_i \geqslant y_j \tag{120}$$

We implemented the ILP in C++ using the ILOG CPLEX Optimization Studio [330] (cf. Chapter 4).

6.3.2.2 *Application to MTIs*

Based on the optimization problem described above, we created a small web application that for a user-defined gene set searches the optimal miRNA set with a size in the range of $\{1, ..., k\}$. For each size $k' \in \{1, ..., k\}$, we solve the implemented ILP to find the best miRNA set of size k' whose combined targetome covers the maximal number of genes in the input. Finally, we list the resulting miRNA set and the corresponding target genes. Additionally, for each analysis, miRPathDB creates an interactive visualization that illustrates the fraction of the input gene list that can be targeted by a particular number of target genes (cf. Figure 77).

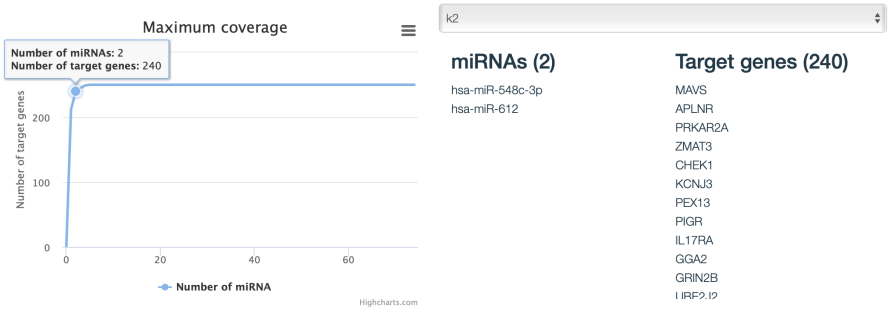


Figure 77: Screenshot of miRPathDB: Results of the maximum coverage ILP for an example gene set.

6.4 DISCUSSION AND CONCLUSION

In this chapter, we presented miRPathDB an intuitive web-based dictionary that provides easy access to information about miRNAs, target genes, and target pathways. Compared to other resources (cf. Figure 78), our database not only focuses on miRNAs from miRBase [281], but also miRNAs and miRNA candidates from its sister project miRCarta [28]. This increases the available data by a ten-fold. Additionally, miRPathDB also provides functionality to compare miRNAs based on sequence, target genes, target pathways, and genomic location. This makes is possible to study similarities and differences of miRNA pairs and to assess associations between the different types of information, e.g., position of the genome, sequence similarity, or overlap of target genes [529]. Although, the main focus of our database is the web interface, the available information can also be accessed via our API (cf. Chapter 4), or downloaded in different text-based formats, such as the GMT format (cf. Section F.2.6). This makes it possible to directly integrate miRNA-pathway links into enrichment analysis tools like GeneTrail [177], GSEA-P [514], or miEAA [30]. Apart from the corresponding data collection, our database also offers interactive analysis tools with which users can study associations between

miRNAs and target pathways from a miRNA- and pathway-centric point of view.
Although miRPathDB is already a valuable resource for the scientific community, there are still some improvements that could be made in future versions of the database. For example, further data sources could be integrated to create biological categories for enrichment analysis. In this context, interesting additions could be marker genes for tissue types, cell types, or even disease phenotypes. Additionally, the support for further species would increase the visibility of our database.
Nevertheless, miRPathDB currently is one of the most comprehensive publicly available resources to study relationships between miRNAs, target genes, and target pathways.

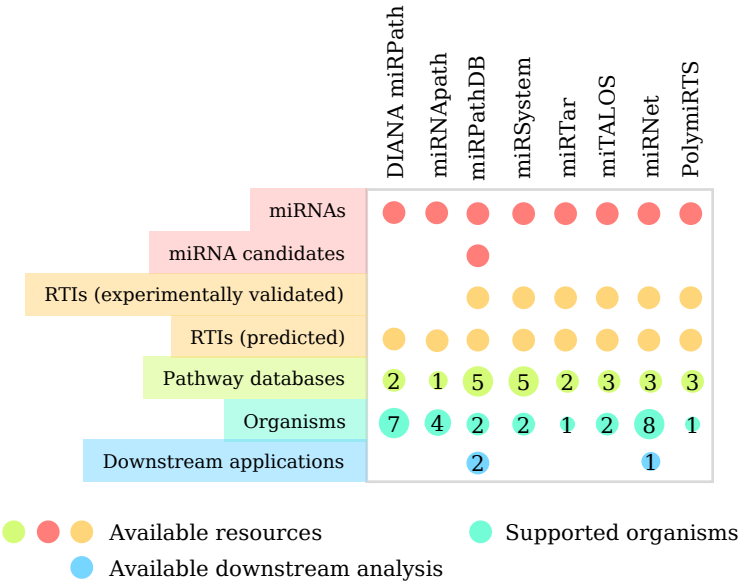


Figure 78: Comparison between miRPathDB and related tools and databases.

7

THE REGGAE ALGORITHM

Author contributions

This chapter is based on our publication "REGGAE: a novel approach for the identification of key transcriptional regulators" [531]. The REGGAE algorithm was mainly developed by Hans-Peter Lenhof and me. The different analyses for this manuscript and their evaluation were conducted by Hans-Peter Lenhof, Lara Schneider, and me. The complete list of contributors can be found in the author list of our publication [531].

In previous chapters, we discussed different approaches for the analysis of deregulated biological processes based on molecular high-throughput profiles. A logical next step is the search for factors that might control or even cause the alterations in these processes. Crucial elements in this context are transcriptional regulators, like transcription factors, co-factors, and chromatin modifiers. These proteins not only have essential functions in most cellular processes [546], but disruptions in their activities have also been observed in a plethora of diseases [293], including heart disease [344, 404], neurodegenerative disorders [381], or cancer [110, 380].
Hence, a crucial task is to assess and evaluate the effects of influential transcriptional regulators in pathological processes. Consequently, a variety of approaches have already been proposed for this purpose. Many of these methods use experimentally validated regulator-target interactions (RTIs), e.g., extracted from ChIP-Seq experiments, to find key regulators. These include (1) ORA-based methods, e.g., TFactS [141] and TED [593], (2) correlation-based methods, like RIF1, RIF2 [437], or CSA [225], and (3) network based methods, such as TFRank [184]. A complete description of all mentioned methods can be found in Section 3.8.1.
In this chapter, we introduce a novel algorithm for the detection of influential transcriptional regulators, called REGGAE (REGulator-Gene Association Enrichment). Our method integrates association measures between regulators and their target genes with a non-parametric enrichment approach to rank key regulators. We demonstrate the capabilities of REGGAE using two application scenarios. The corresponding results prove that our algorithm has a superior performance compared to the approaches listed above and constitutes a powerful tool for uncovering complex regulatory mechanisms.

7.1 ALGORITHM

We designed the REGGAE (REGulator-Gene Association Enrichment) algorithm to identify and prioritize transcriptional regulators that strongly influence the expression of a given gene list, i.e., the most deregulated genes identified in a group comparison.
Our algorithm depends on two types of inputs. First, a matrix containing normalized gene expression measurements for n samples that are assigned to different sample groups, such as disease and control. Additionally, REGGAE requires a set of experimentally validated regulator binding sites (i.e., RTIs, cf. Section 3.2.4.1). Based on these inputs, the following processing steps are conducted to identify the most influential regulators.

7.1.1 *Group comparison and feature selection*

The first step of each REGGAE analysis is comparing gene expression values between the two given sample groups. For this step, any statistical measure described in Section 3.4 can be used, e.g., log-fold-changes.
We then select the most up- or down-regulated genes from the resulting score list, i.e., the genes for which we want to identify the most influential regulators. To this end, our framework offers different strategies. Users can select (i) the number of genes that should be used, (ii) all genes with scores above or below a certain threshold, or (iii) genes that are contained within a specified quantile of the ranked list, e.g., the upper 10% quantile. An overview of this step is depicted in Figure 79.
While REGGAE can be applied to assess the impact of regulators for both the most up- and downregulated genes, we restrict the description of the subsequent paragraphs to up-regulated ones. We denote the respective gene set as $D = \{g_1, g_2, ..., g_m\}$. Finally, we sort D with respect to their gene score.

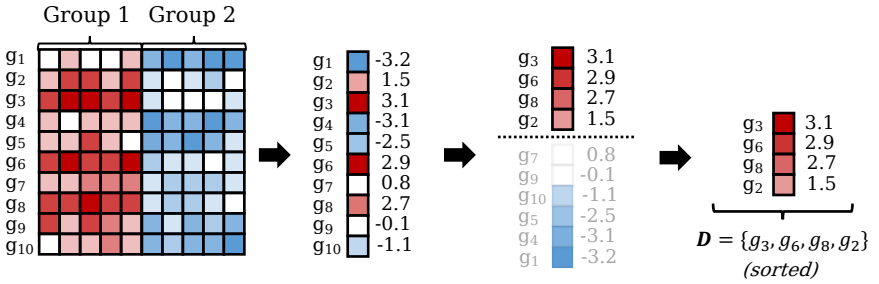


Figure 79: Overview of the group comparison and feature selection step of the REGGAE algorithm. First, the expression differences between the two sample groups are calculated. Then, the most up-regulated features are selected.

7.1.2 *Association analysis*

For each gene $g_i \in D$, the next step in our workflow is to evaluate which regulators might have the biggest influence on g_i. To this end, we calculate an association score between g_i and each regulator from our RTI collection that targets g_i. Subsequently, we sort the regulators according to their (absolute) association with g_i. We denote the resulting list $R_{g_i} = \{r_{i1}, r_{i2}, ...\}$. An overview of this procedure is shown in Figure 80.
Our framework offers different statistical measures to calculate the association between genes and regulators. These include Pearson's correlation coefficient [408], Spearman's rank correlation coefficient [504], and distance correlation [279] (cf. Section 3.5).

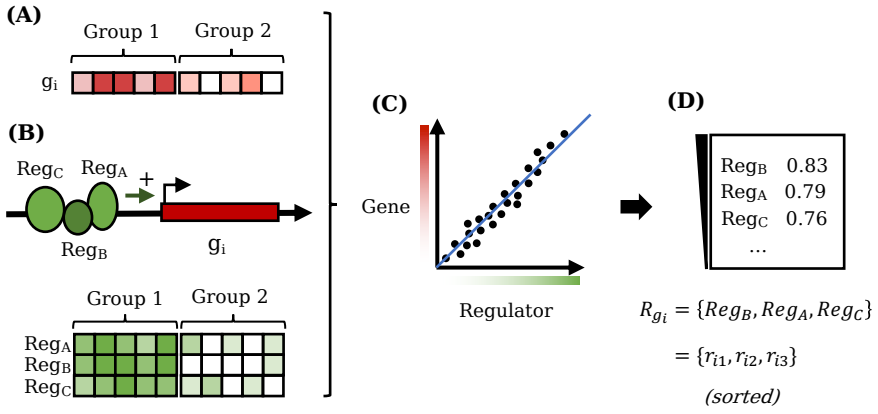


Figure 80: Overview of the association analysis step of the REGGAE algorithm. (A) Given expression values for gene g_i and (B) corresponding expression values for all regulators from our RTI database that target g_i: Reg_A, Reg_B, and Reg_C. (C) We can then calculate an association measure to assess the strength of their relationship. (D) We denote the resulting (sorted) list $R_{g_i} = \{r_{i1}, r_{i2}, r_{i3}\}$.

7.1.3 *Data integration*

Our final goal is to assess the total impact of each regulator. For this purpose, we have to combine the list of deregulated genes D with the results of the association analysis, while retaining the order in the gene and regulator lists. To this end, we use a non-parametric encoding that creates a new ordered list of regulators, where each regulator is contained up to m times depending on the number of target genes in D. This list can then be used in an enrichment analysis to assess the overall impact of each regulator (cf. Section 7.1.4).

Given the sorted gene list $D = \{g_1, ..., g_m\}$ and for each gene g_i the corresponding regulators $R_{g_i} = \{r_{i1}, r_{i2}, ...\}$, sorted according to their association score, we can now define a new List $L = \{r_{11}, r_{21}, ..., r_{m1}, r_{12}, r_{22}, ...\}$ that contains all r_{ij} ordered in a column-wise fashion. For each gene in D, we first add the regulators with the highest association, then the ones with the next highest association and so on. An overview of this approach is shown in Figure 81.

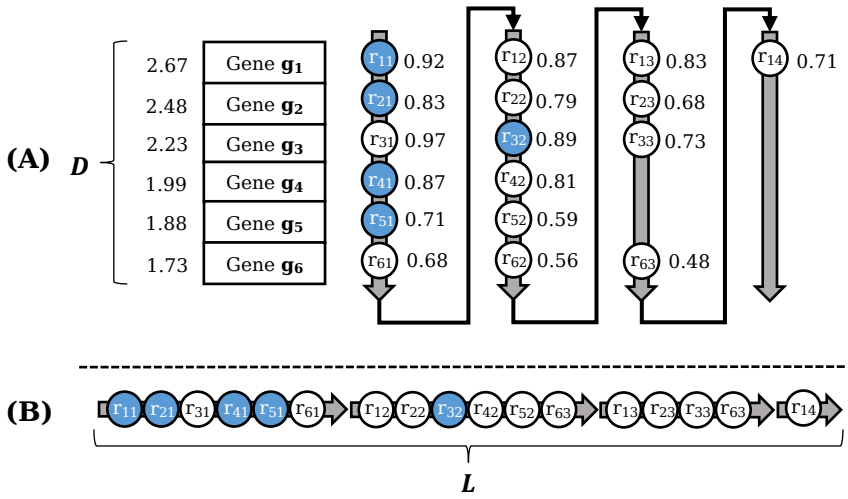


Figure 81: Data integration step of the REGGAE algorithm. (A) Given the gene list $D = \{g_1, ..., g_6\}$, which is sorted with respect to the degree of deregulation of each gene, and for each g_i a regulator list that is sorted according to their influence on g_i. The blue circles represent one specific regulator with five target genes in D. (B) The new list L is created by sorting the regulators in a column-wise fashion. This figure was adapted from [531].

7.1.4 *Enrichment analysis*

After the data integration step, we obtain an ordered list L of regulators. In this list, each regulator with m target genes is contained up to m times in the list of deregulated genes (D), for example, the blue regulator in Figure 81B has five targets.
We assume that regulators with a strong influence on their respective target genes in D are enriched at the beginning of L, i.e., that for each target $g_i \in D$ it was one of the regulators with the highest association score in the respective regulator list R_{g_i}.
In order to test this hypothesis, we conduct a non-parametric enrichment analysis for each regulator. To this end, our framework offers two methods: (i) the Kolmogorov-Smirnov (KS) test and (ii) the Wilcoxon rank-sum (WRS) test (cf. Section 3.4 + 3.6). This is depicted in Figure 82.

Finally, the p-values for all regulators are adjusted to account for the multiple testing problem. This can be done by any of the methods described in Section 3.3.2.

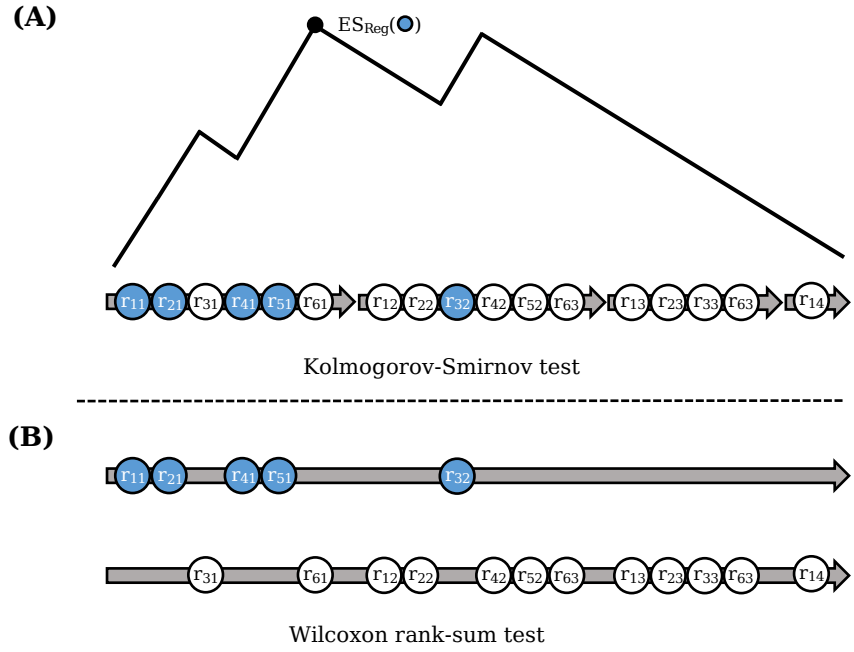


Figure 82: Enrichment analysis step of the REGGAE algorithm. Illustration of the two available enrichment analysis procedures (A) Kolmogorov-Smirnov test and (B) Wilcoxon rank-sum test. The blue circles represent one specific regulator with five target genes. This figure was adapted from [531].

7.1.5 *Bootstrapping*

High-throughput measurements can contain technical noise that might affect the different processing steps of REGGAE and, consequently, impede the final regulator ranking. In order to account for potential biases in our results and to increase the general robustness of our method, we have developed a bootstrapping [135] approach that can optionally be applied.
This resampling procedure includes all processing steps described in the previous sections except group comparison and feature selection, i.e., the list of deregulated genes D remains constant.
All remaining processing steps are repeated $B \in [1,000, 10,000]$ times using the following bootstrapping strategy [531]:

1. Create a resampled matrix E by selecting n random samples with replacements from the original gene expression matrix.

2. Repeat the association analysis, data integration, and enrichment analyses steps using E as input.

In the end, we obtain one result list for each of the B bootstrapping runs. These lists are then combined to a final result. To this end, we assign the median unadjusted p-value of all B replicated results to each regulator. The final p-values for all regulators are then adjusted as described above. Additionally, we can use the different replications to calculate standard deviations and confidence intervals [134] that can help to evaluate the significance and robustness of the results.

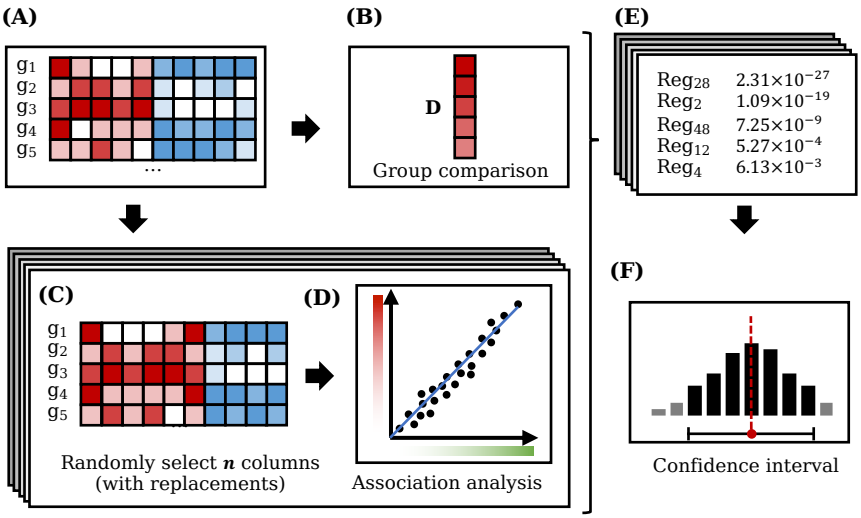


Figure 83: Bootstrapping replication step of the REGGAE algorithm. (A) The original gene expression matrix is used to calculate (B) the most deregulated genes. This list remains unchanged. (C) For the remaining steps, we use bootstrapping to create $B \in [1,000, 10,000]$ resampled gene expression matrices. (D+E) Each of these matrices is then used to conduct all remaining processing steps of the REGGAE algorithm. (F) The replicated results can then be used to obtain more stable p-values, standard deviations, and confidence intervals.

7.1.6 *Aggregating REGGAE results*

In addition to the bootstrapping strategy described in the last paragraph, we also offer the possibility to combine the results of several REGGAE analyses, e.g., for a different number of deregulated genes. This can also help to improve the final regulator ranking. The different results can either be aggregated via their rank or via their p-value. For ranks, we use the sum of all ranks, and for p-values we either use the maximum or the second-order statistic [123].

7.2 RESULTS

In the following sections, we present and compare the results of REGGAE and the alternative approaches described in Section 3.8. To this end, we applied all methods in two distinct application scenarios.
First, we analyzed breast cancer cell lines to identify transcription factors and chromatin modifying proteins that could be responsible for expression changes between two clinically relevant subtypes, i.e., tumor cells that express the estrogen receptor on their surface (ER-positive) and tumor cells that do not express them (ER-negative).
In the second scenario, we applied all methods to expression profiles in which the activity of one specific transcription factor is perturbed, i.e., artificially induced or repressed. We then evaluated which approaches are able to successfully detect the influence of the disrupted transcription factors.

7.2.1 *Comparison of ER-positive and ER-negative breast cancer cells*

Breast cancer is, with around 30% of all diagnosed cancer cases, the most abundant cancer type among women [489]. One of the most important clinical marker for breast cancer is the availability of hormone receptors on the surface of cells, especially the estrogen receptor (ER). ER-positive (ER+) tumors make up around 70% of all cases [151] and seem to have a better prognosis than ER-negative (ER-) tumors [34]. In the following, we analyzed a data set of 37 breast cancer cell lines [385] to identify transcriptional regulators that may cause expression differences between ER+ and ER- tumors.

7.2.1.1 *Data set and processing steps*

The data set by Neve et al. contains gene expression profiles of 16 ER+ and 21 ER- breast cancer cell lines. To calculate expression differences between the two classes, we applied the Shrinkage t-test (cf. Section 3.4.3.2). From the sorted result list, we created five different test sets. First, we selected the 250, 500, 750, and 1000 genes with the highest t-score. Additionally, we created a gene list with the most significantly up-regulated genes ($P < 0.01$) in the ER+ group, i.e., 1719 genes. We then applied REGGAE and the competing methods to all five test sets. Finally, we aggregated the five result lists for each method. For the for methods that calculate p-values, i.e., CSA, REGGAE, TED, and TFactS, we used the largest of the five p-values for each regulator. For the remaining methods, i.e., RIF1, RIF2, TDD, and TFRank, we applied rank aggregation. In order to compare the final result lists, we use all significant results for all approaches that calculate p-values and the 200 most highly ranked regulators for the remaining ones. The complete set of parameters for all methods can be found in Appendix E.2.

7.2.1.2 *Robustness of REGGAE*

First, we analyzed the robustness of REGGAE with respect to the number of bootstrap replications and test set size. In order to find the number of bootstrap replications needed to obtain stable results, we applied our method to each of the five test sets and saved the results after each iteration of the bootstrapping approach. In order to assess the robustness, we calculated the sum of regulator pairs that swap positions after each iteration. Given two regulators a and b and their ranks in the sorted result list after iteration i: $r_i(a)$ and $r_i(b)$. We say a and b change their order iff $r_{i-1}(a) < r_{i-1}(b)$ and $r_i(a) > r_i(b)$ or vice versa.
The results for the test set with $1,000$ genes are depicted in Figure 84A. As illustrated, the number of fluctuating regulator pairs rapidly decreases with increasing number of iterations and converges after approximately $1,000$.

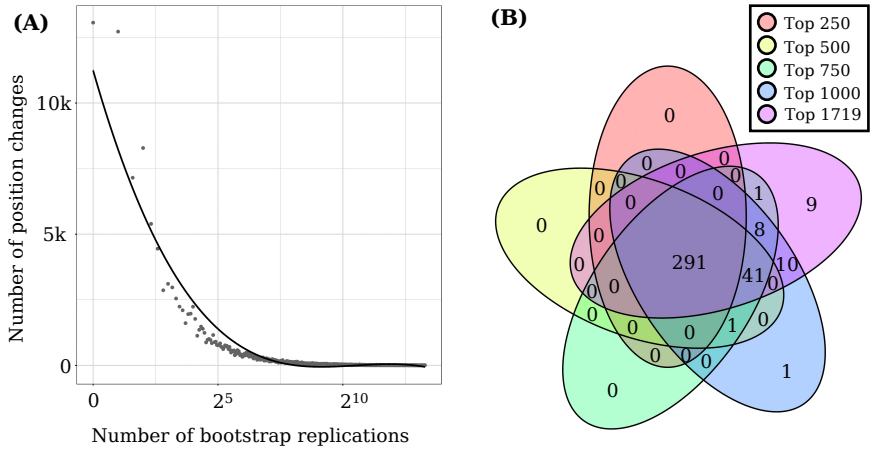


Figure 84: Robustness of the REGGAE algorithm. (A) Scatterplot depicting the dependency between the number of bootstrap replications and the position changes in the REGGAE result list for the bootstrapped input matrix and the test set of length 1000. (B) Overlap of the REGGAE results for the test sets of different lengths. This figure was adapted from [531].

We also tested how the length of the input list affects the REGGAE results. For this purpose, we analyzed the overlap of significant regulators for all five gene lists. The corresponding Venn diagram is shown in Figure 84B. As depicted in this plot, the resulting set of regulators remains highly stable, although the number of significant regulators increases with larger test set sizes. We observe the largest increase (42 new regulators) when 500 genes are considered instead of 250. The analysis of even longer lists only identified a few additional significant regulators.

7.2.1.3 *General comparison of* REGGAE *and alternative approaches*

We also used the breast cancer data set by Neve et al. to evaluate the performance of REGGAE in comparison to the approaches discussed in Section 3.8.
Since many of the used approaches employ statistical tests that in part have very distinct null hypotheses, the results of all analyses need to be interpreted with utmost caution.
Nevertheless, the following analyses assess if the different approaches can detect relevant regulators that might contribute to expression differences between ER+ and ER- cancer cells. First, we analyzed if the different approaches produce similar results, i.e., if the identified regulators have a significant overlap. To this end, we compare the aggregated result lists using the hypergeometric test (cf. Section 3.6.2). The corresponding Venn diagrams are depicted in Figure 85.
Five of the seven tested algorithms (CSA, RIF1, RIF2, TFactS, and TFRank) have a highly significant overlap with the REGGAE result list. The remaining methods (TED and TDD) have minimal overlap. REGGAE and TED even have no joint results in this analysis.

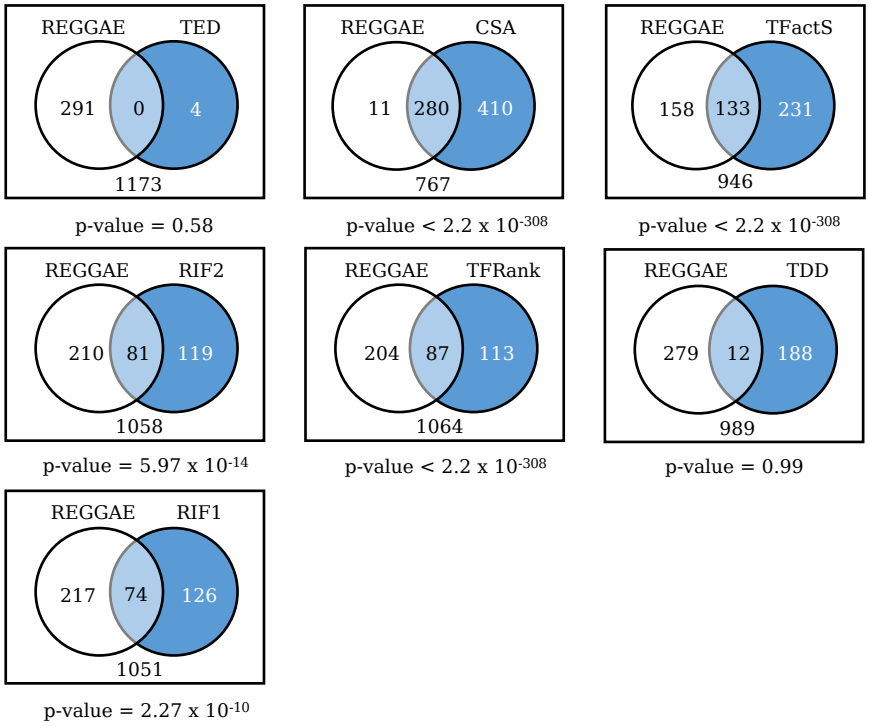


Figure 85: Overlap between the aggregated result lists of REGGAE and all other methods. P-values were calculated using the hypergeometric distribution. This figure was adapted from [531].

Although the results of most methods have a significant overlap with REGGAE, the actual rankings of the regulators in the result lists are different. The five most significant regulators identified by REGGAE are depicted in the first column of Table 4. The remaining columns of Table 4 and 5 show if other algorithms have also detected the regulators and at which position in the ranked list. In general, we observe that CSA and TFRank also detect all five regulators. TFRank even identifies all five amongst the most important ones, which is not the case for the other methods. RIF1 and RIF2 only identified four of the five genes among its top 200 results, TFactS only two, and TDD and TED none.

Regulator	**REGGAE**	**CSA**	**TED**	**TFactS**
FOXA1	6.34×10^{-141}(1)	9.76×10^{-6}(359)	1.0(843)	1.0(953)
GATA3	3.23×10^{-137}(2)	9.76×10^{-6}(421)	1.0(681)	0.05(369)
ESR1	6.52×10^{-129}(3)	9.76×10^{-6}(509)	1.0(440)	1.0(790)
MYB	6.34×10^{-125}(4)	9.76×10^{-6}(262)	1.0(6)	0.31(519)
SPDEF	2.60×10^{-118}(5)	9.76×10^{-6}(40)	1.0(892)	3.6×10^{-19}(32)

Table 4: Comparison of the top 5 candidates in the REGGAE result list and alternative approaches (Part 1): CSA, TED, and TFactS. For all methods, FDR-adjusted p-values are shown. Additionally, the rank of each regulator in the sorted result list of each method is depicted in parentheses. This figure was adapted from [531].

Regulator	**RIF1**	**RIF2**	**TDD**	**TFRank**
FOXA1	−2.87(116)	8.34(18)	8.4×10^{-6}(956)	6.92(2)
GATA3	−2.73(113)	5.16(62)	8.7×10^{-6}(747)	6.56(3)
ESR1	−1.93(229)	−0.10(915)	8.4×10^{-6}(949)	10.28(1)
MYB	−2.07(130)	4.14(75)	8.4×10^{-6}(878)	5.45(6)
SPDEF	−3.05(32)	8.54(15)	1.4×10^{-5}(434)	6.44(4)

Table 5: Comparison of the top 5 candidates in the REGGAE result list and alternative approaches (Part 2): RIF1, RIF2, TDD, and TFRank. For all methods, the test statistics are shown. Additionally, The rank of each regulator in the sorted result list of each method is depicted in parentheses. This figure was adapted from [531].

7.2.1.4 REGGAE *identified key regulators of estrogen signaling*

Next, we investigated if the five most significant regulators in the aggregated REGGAE result list (ESR1, FOXA1, GATA3, MYB, and SPDEF) can help to explain the gene expression differences between ER+ and ER- breast cancer cells and might even be involved in the increased malignancy of the ER- group.
Most notably, we observe that one of the two human estrogen receptors ESR1, was identified as one of the most important factors in our analysis. Since the availability of ESR1 on the surface of the tumor cells is amongst the key distinguishing features of the two analyzed groups, this is a strong indicator that REGGAE produces biologically relevant results.
Furthermore, a close inspection of each candidate revealed that all of the five regulators are well-known prognostic markers of breast cancer and that their expression is directly associated with a more favorable outcome of the disease [351, 352, 545, 574]. ESR1, FOXA1, GATA3, and MYB are also regularly mutated in breast cancer [273].
Moreover, we found that the different regulators are involved in several relevant biological processes in breast cancer and often work together. For example, FOXA1, ESR1, and GATA3 have been described to be co-expressed [457] and co-localized [275] in breast cancer cells. This indicates that these proteins could interact or might even belong to the same regulator complex. Indeed, Kong et al. present evidence that they are part of an enhanceosome that directly controls the estrogen receptor signaling cascade [275]. Moreover, ESR1, FOXA1, GATA3, and SPDEF have been discussed as master regulators in the fibroblast growth factor receptor 2 (FGFR2) signaling pathway, which is strongly related to breast cancer risk [156].

7.2.2 *Perturbation signatures*

One interesting way to study the effects of a transcriptional regulator is to analyze gene expression profiles in which the activity of a certain regulator is artificially perturbed. Different kinds of perturbations can be used to simulate specific genetic or molecular events, e.g., a regulator's knock-out can mimic loss-of-function mutations, while induced over-expression of a regulator can simulate activating modifications.
In this section, we present results based on four gene expression data sets that contain perturbed samples and corresponding unaffected controls. In the first data set the expression of the transcription factor MYC is artificially overexpressed and in the remaining three data sets one particular regulator is knocked out, i.e., NANOG, POU5F1, and SOX2. In the following, we applied REGGAE and the competing

approaches to each of the four data sets in order to evaluated if the different methods were able to detect the perturbed regulator.

7.2.2.1 *Overexpression of* MYC *in mouse lymphoma cells*

First, we analyzed a mouse lymphoma data set with artificially induced overexpression of the MYC proto-oncogene. MYC is an important regulator of cell growth, proliferation, and pluripotency [109]. The overexpression of MYC is often associated with the malignancy of the disease. In particular, B cell lymphomas are characterized by high levels of MYC [490]. Eμ − Myc transgenic mice contain a genetic alteration that resembles MYC activation in B cells [70, 205]. Hence, it is often used to study MYC-controlled lymphoma development.
Here, we compared gene expression profiles of 50 B cell lymphoma samples from Eμ − Myc transgenic mice and ten samples of lymph nodes from healthy wild-type mice [369]. As in the first application scenario, we used the Shrinkage t-test to compare the expression between the two sample groups (cf. Section 3.4.3.2). From the resulting list, we then created two different test sets: the 250 most up-regulated and the 250 most down-regulated genes. Finally, we applied the different approaches to find the most influential regulators for both lists. All processing steps and parameters are described in Section E.5.
The results of all algorithms are shown in column (A) of Table 6. As depicted, CSA, REGGAE, RIF1, and TFRank successfully identified MYC as important regulators for both up- and down-regulated genes. RIF2 and TFactS detected the proto-oncogene only for the test set with up-regulated genes. TDD and TED could not retrace the effect of MYC. While most methods, were able to detect MYC, only REGGAE and TFRank could successfully rank the gene amongst the top candidates in both lists.

Additional regulators identified by REGGAE

Apart from the proto-oncogene itself, REGGAE was able to detect many regulators directly controlled by MYC (cf. Supplementary Table 38). In particular, REGGAE identified many chromatin modifiers, for example KAT2A, SMC3, SUZ12, RCOR, SMC1A, SCMARCA4, SCMARCA5. This emphasizes the well-established role of MYC in chromatin remodeling in general [272] and especially in B cells [266]. Additionally, the list also contains several genes directly involved in MYC signaling, such as E2F1 and E2F4 [295, 434], and also MYC hallmark genes, like RAD23B and TRIM28.
In summary, the above results further emphasize the ability of our algorithm not only to detect important regulators directly but also downstream effects.

7.2.2.2 *Knock-outs of* NANOG, POU5F1, *and* SOX2

Pluripotency factors, such as NANOG, POU5F1 (OCT4), or SOX2, are essential transcription factors in human embryonic development. As the name suggests, these elements control mechanisms that maintain pluripotency in human embryonic stem cells (ESCs), but are also involved in cell fate determination or self-renewal of these cells [313]. In this section, we analyzed a data set from Wang et al. [561] that contains gene expression profiles of normal ESCs and samples with knock-outs of NANOG, POU5F1, or SOX2. For each pluripotency factor, we compared knock-out and control samples to evaluate if all approaches can retrace the effects of the respective regulator.

	(A)	**(B)**		
Method	**MYC**	**NANOG**	**POU5F1**	**SOX2**
CSA	**281** | **126**	**574** | 571	510 | **273**	510 | **259**
REGGAE	**1** | **1**	**1** | **91**	**1** | **1**	**6** | **4**
RIF1	**126** | **186**	791 | **148**	795 | **171**	**285** | **555**
RIF2	**8** | 251	**114** | **193**	762 | **190**	332 | **34**
TDD	466 | 492	815 | 771	822 | 800	467 | 523
TED	208 | 225	567 | 501	683 | 588	682 | 682
TFactS	**404** | 528	**318** | **308**	**531** | **319**	**170** | **99**
TFRank	**1** | **3**	**113** | **2**	**200** | **1**	499 | **1**

Table 6: Comparison of REGGAE and the alternative methods based on different perturbation signatures: (A) Over-expression of MYC, and (B) knock-outs of NANOG, POU5F1, and SOX2. The numbers depict the rank of the analyzed regulator in the respective result list. Blue numbers indicate if the corresponding result is considered significant. This figure was adapted from [531].

In order to calculate gene expression differences between knock-out experiments and controls, we use the Shrinkage t-test (cf. Section 3.4.3.2). From the resulting lists, we picked both the 250 genes with the highest and lowest scores. We then applied all algorithms to both test sets to find the most relevant regulators. All processing steps and parameters are described in Section E.6. A summary of the results is shown in Table 6B.

As depicted, REGGAE and TFactS are the only methods that could identify the perturbed transcription factor in all six cases. Of the remaining methods, TFRank detected five correctly, RIF1 four, CSA three, and RIF2 two. As in our MYC perturbation example, REGGAE and TFRank are by far the best methods in terms of ranking.

REGGAE found the perturbed transcription factor in five out of six cases amongst the top ranked results and TFRank in three.

7.3 DISCUSSION AND CONCLUSION

In this chapter, we presented REGGAE, a novel method for the identification of influential regulators that may be involved in deregulated processes. Our algorithm incorporates association scores between regulators and corresponding targets into an enrichment-based scoring scheme that is used to rank regulators. The combination of the two approaches significantly improves the prioritization of the regulators and set REGGAE apart from competing algorithms.
Although REGGAE was able to outperform alternative methods in the described application scenarios, there are still some extensions or modifications of the approach that could potentially further improve the results.
One of these extension is the utilization of additional association measures. Currently, we mainly use Pearson's and Spearman's correlation coefficients (cf. Section 3.5), as both measure the relationship between regulators and target genes as well as the direction of the relationship. Additionally, we also implemented the distance correlation that can be applied to study arbitrary relationships, but has no information about the association direction. Apart from these methods, alternative measures could potentially be applied, such as entropy based methods like mutual information, maximum information coefficient (MIC) [223], or the universal dependency score (UDS) [389].
Moreover, all approaches discussed in this chapter rely on experimentally validated RTIs. Most RTIs are defined by ChIP-Seq peaks of a specific regulator in a predefined window around the TSS. While this covers the promoter region and depending on the window size proximal enhancers, information about distal regulatory elements is currently missing. Although it has been shown that enhancers often have the strongest effect on the nearest genes [139], the inclusion of additional distal regulatory regions, e.g., enhancer regions from GeneHancer [154] might further improve the results. While the biggest advantage of using RTIs is that these are supported by experimental evidence, there are also some disadvantages. In particular, that experimental binding information is often only available for certain cell types. Hence, in the best case the respective interactions should only be used to analyze data sets with the same cell type. However, currently there are not enough data sets available to create such cell type specific analyses. To overcome this problem, we had to combine the information for different cell types to create our RTI collection. Consequently, this can lead to false positive or false negative interactions. However, we assume that a small number of false RTIs only has moderate influence on the REGGAE results.

Alternatively, it might also be possible to use predicted binding sites instead of experimentally validated ones, e.g., produces by FIMO [190]. This approach seems to deliver relative accurate results if binding site predictions are combined with epigenetic signals, like open-chromatin regions [412, 472]. However, in this case further experimental data sets would be required.
Moreover, all methods discussed in this thesis currently analyze the effect of regulators individually. However, regulators are often part of protein families with similar function or form large complexes that have a joint influence on the expression of specific target genes. Hence, it might also be beneficial to incorporate information about protein families or complexes into future analyses.
Nevertheless, the analyses conducted in this chapter show that REGGAE is already well equipped for the analysis of influential regulators and outperforms most of the competing methods. While most approaches detect overlapping feature sets, REGGAE and TFRank produce by far the best rankings. Furthermore, REGGAE stands out by providing confidence measures, like confidence intervals and p-values, that help to evaluate the robustness and significance of the results. On top of this, the utilized association measure can be used to estimate if a specific regulator has an activating or repressing effect. In summary, our results demonstrate that REGGAE is a powerful tool for the elucidation of deregulated regulatory mechanisms.

8

THE REGULATORTRAIL WEB SERVICE

Author contributions

This chapter is based on our publication "RegulatorTrail: a web service for the identification of key transcriptional regulators" [533]. The RTI-based workflows described in this chapter were designed by Hans-Peter Lenhof and me. The Motif-based workflows were designed by Florian Schmidt and Marcel Schulz. The web service was mainly implemented by me with further contributions by Lara Schneider and Florian Schmidt. The complete list of contributors can be found in the author list of the manuscript [533].

Due to the inherit importance of transcription factors and chromatin modifying proteins in the control of nearly all biological processes, many approaches have been developed to study their effects based on molecular high-throughput profiles.
One essential class of approaches are the RTI-based methods discussed in the last chapter (cf. Chapter 7). These use experimentally validated binding sites of regulators to assess their influence on a given set of target genes. In general, RTI-based methods can be divided into different categories that require different input types and that employ different analysis strategies.
The first category are ORA-based approaches, like TED [593] and TFactS [141]. The input for these methods is an unordered set of input genes that is used study significant overlaps with regulator targets. The next category are network-based approach, such as TFRank [184], that interpret the RTIs as a directed graph, where nodes constitute regulators and target genes, and each edge a specific RTI. In addition to this graph, TFRank requires importance scores for each analyzed target gene that are used as vertex weights in the graph. These weights are then propagated through the network to find the most important regulators (cf. Section 3.8.1.4). The last category are correlation-based approaches, like CSA [225], REGGAE [531], RIF1, and RIF2 [437]. The input for these methods is a expression matrices with multiple samples that belong to two groups, e.g., disease vs. control. The expression measurements are then employed to assess the relationship between regulators and target genes.
Another class of methods predicts the binding patterns of transcription factors based on the DNA sequence and epigenetic modification patterns [197, 412], e.g., BinDNase [249], CENTIPEDE [412], HINT-BC [196], MILLIPEDE [321], PIQ [485], or TEPIC [472]. The input for these methods are DNA binding motifs and epigenetic measure-

ments, such as open-chromatin regions. First, these measurements are used to identify genomic regions that are accessible to DNA binding proteins. The obtained regions are then scanned using the DNA binding motifs to identify potential regulator binding sites. The resulting predictions are then often used in downstream applications. For example, the PASTAA web service conducts enrichment analyses for the targets of each regulator to identify potentially affected biological categories [448]. Many authors also used the regulator binding predictions as features in machine learning applications, such as predictive models of gene expression [80, 103, 347, 377, 472].
Most of the methods described above are available as stand-alone applications that need to be installed or even R-packages that require programming experience. Additionally, some approaches are also available as web services, e.g., Enrichr [284], PASTAA [448], Regulatory Snapshots [183], or TFactS [141]. While these web services are more accessible to non-expert users, they are often restricted to one application scenario. Hence, a user friendly solution is required that allows both expert and non-expert users to conduct powerful analyses of influential regulators in different application scenarios.

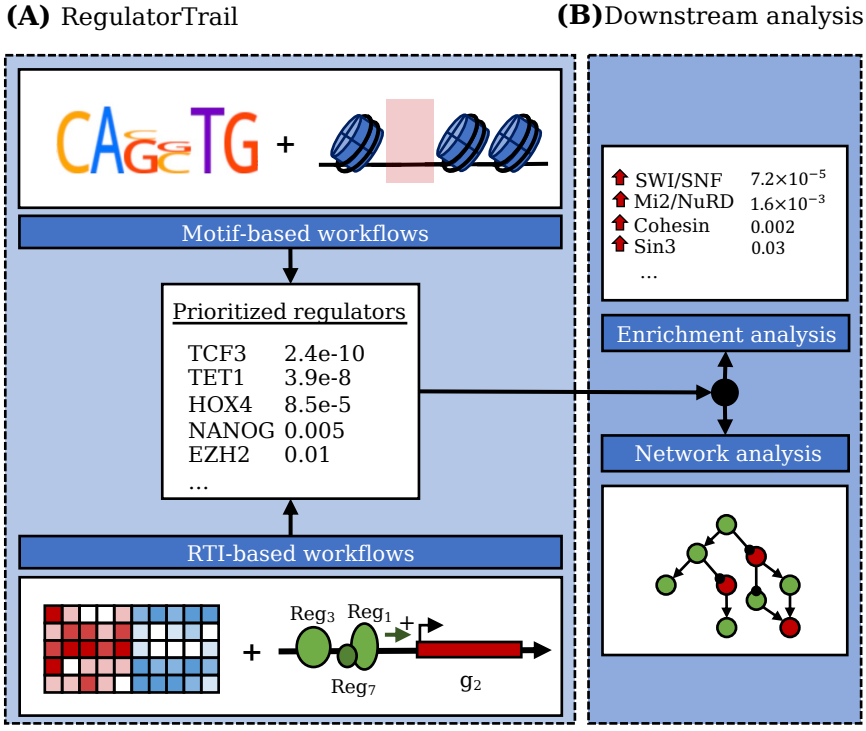


Figure 86: General overview of RegulatorTrail. (A) Our web service offers both motif-based and RTI-based approaches for the identification of key regulators. (B) The result list of all workflows can be used in a downstream enrichment or network analysis. This figure has been adapted from [533]

8.1 THE RegulatorTrail WEB SERVICE

Here, we present the RegulatorTrail web service, a comprehensive toolbox for the analysis of transcription factors, co-factors, and chromatin modifiers. Our web service offers various methods for the identification and prioritization of influential regulators that combine binding information of regulators with transcriptomic, proteomic, or epigenomic data sets. In contrast to other approaches that focus on one specific task, we designed RegulatorTrail as a general framework for the analysis of influential regulators in a broad range of application scenarios and a diverse set of input types. In particular, we ensured that the provided methods cover at least one approach from each of the categories mentioned above.
In total, our web service offers seven RTI-based approaches and two motif-based methods that can be applied to analyze data sets for five different organisms: *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*, and *Rattus norvegicus*. The result of all analyses is a sorted list of transcriptional regulators. These lists can either be downloaded by the user, visualized and inspected in the browser, or even directly be used in a downstream enrichment or network analysis (cf. Figure 86B). In the following sections, we describe the different workflows of our web service in more detail.

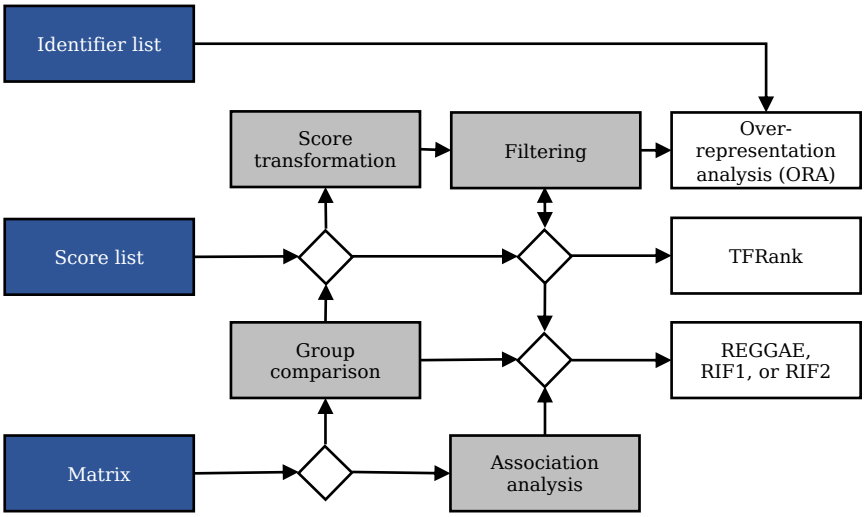


Figure 87: Overview of the RTI-based RegulatorTrail workflows.

8.1.1 *RTI-based workflows*

RegulatorTrail offers seven methods for the analysis of influential regulators based on our large collection of experimentally validated RTIs (cf. Section 3.2.4.1). These can be applied in three distinct application scenarios that require different input types to identify key regulators. An overview of all RTI-based workflows of our framework is depicted in Figure 87.

8.1.1.1 *Scenario 1: Regulator target over-representation analysis (ORA)*

The first class of approaches are ORA-based methods that analyze for all regulators in the considered RTI collection if their respective target genes have a significant overlap with a given test set. Consequently, the input of this scenario is a set of gene or protein identifiers and a corresponding reference set, e.g., all protein coding genes. For the ORA-based identification of the most relevant transcriptional regulators, our framework offers three statistical tests: the binomial test as proposed by Yang et al. [593], the hypergeometric test as presented by Essaghir et al. [141], and Fisher's exact test (cf. Section 3.8.1.3). Since one statistical test is applied for each regulator, we run into the multiple-testing problem and all resulting p-values need to be adjusted. To this end, any method described in Section 3.3.2 can be applied. An overview of the described approach is depicted in Figure 88A.

8.1.1.2 *Scenario 2: Network-based analysis*

The input for the second scenario is a gene list with associated scores, e.g., the most deregulated genes in a group comparison. In contrast to Scenario 1, where the input is an unordered score list, here, each score indicates the importance of a gene, e.g., the degree of deregulation.
In addition to ORA-based methods that can also be applied, RegulatorTrail also offers a network-based heat diffusion approach that is based on TFRank (cf. Section 3.8.1.4, [184]). For this purpose, the RTIs are interpreted as a directed graph, where vertices represent regulators and target genes and edges interactions between them. The scores from the uploaded test set are then used as initial vertex weights. To find the most influential regulators, the heat diffusion algorithm described in Section 3.8.1.4 is applied to propagate the initial weights through the inverted network, i.e., from target genes to the associated regulators. Finally, all regulators are sorted based on the assigned scores. This approach is illustrated in Figure 88B.

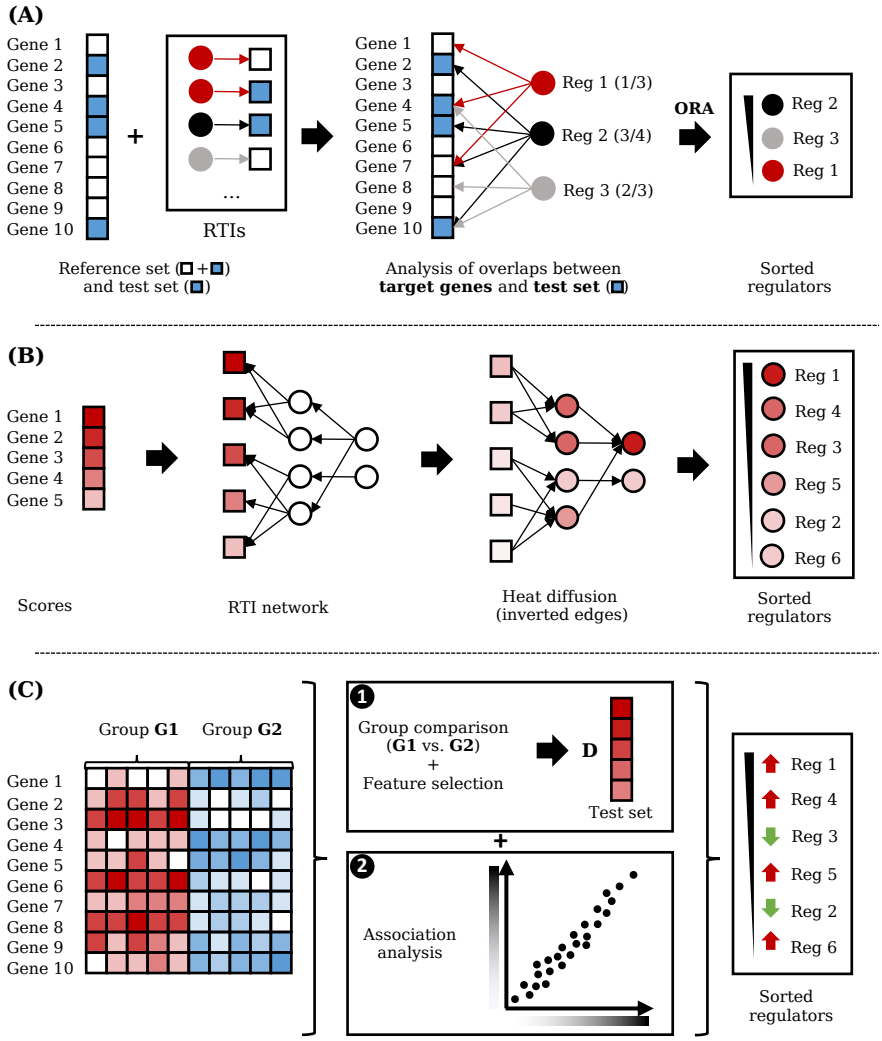


Figure 88: RTI-based application scenarios: (A) ORA-based analysis for gene sets, (B) Network-based analysis for score lists, and (C) Association-based analysis for expression matrices.

8.1.1.3 *Scenario 3: Association-based analysis*

The input for the third application scenario is a matrix with gene expression measurements from two groups of samples, e.g., disease vs. control. The availability of multiple data points per gene makes it possible to estimate the association between regulators and their target genes.

Hence, in addition to the methods described in Scenario 1 and 2, RegulatorTrail provides further approaches that can be applied if gene expression matrices are uploaded by a user: REGGAE (cf. Chapter 7, [531]), and the regulatory impact factors RIF1 and RIF2 (cf. Section 3.8.1.2, [437]). All three methods combine both gene expression differences between the two groups (cf. Section 3.4) and association

measures, e.g., correlation coefficients, to prioritize all regulators. An overview of this scenario is depicted in Figure 88C.

8.1.2 *Example applications of RTI-based workflows*

Applications of RTI-based methods have already been discussed as part of the REGGAE chapter of this thesis (cf. Chapter 7). Apart from these analyses, we also used REGGAE to conduct a comprehensive analysis of a Wilms' tumor (WT) data set that revealed several regulatory mechanisms, which may contribute to the elevated malignancy of a particular WT subtype. The results of this study are presented in Chapter 9.

8.1.3 *Motif-based workflows*

RegulatorTrail also offers two workflows for the analysis of transcriptional regulators based on our collection of DNA binding motifs (cf. Section 3.2.4.2). For this purpose, we utilized the TEPIC framework developed by Schmidt et al. that provides implementations for the functions described in the following paragraphs [472, 473].

8.1.3.1 *Scenario 4: Estimating transcription factor binding affinities using* TEPIC

The inputs for the fourth scenario are open-chromatin regions in standard BED format (cf. Section F.2.1), i.e., genomic regions that are accessible to DNA binding proteins. These can, for example, be obtained from DNase-seq (cf. Section 3.1.1.2) or histone ChIP-seq (cf. Section 3.1.1.2) experiments.
The uploaded regions are then used to calculate binding affinity scores for each gene and all considered motifs (cf. Figure 89). To this end, our web service applies the following method for each gene individually. First, we select all open-chromatin regions that overlap with a user-defined window around the transcription start site of the considered gene. In the next step, the DNA sequence of the resulting genomic regions are scanned using DNA binding motifs to find potential regulator binding sites. To this end, our framework applies TEPIC's biophysical model with exponential decay function to estimate binding affinity scores of each regulator in our collection (cf. Section 3.8.2.2).
The resulting affinity scores can then be downloaded by the user or processed in two different ways. On the one hand, the affinity scores of one particular regulator and corresponding target genes can be used in an enrichment analysis to detect biological processes that might be affected by this regulator. On the other hand, the affinity scores can be combined with gene expression data to find the regulators that might explain these expression values (cf. Section 8.1.3.2).

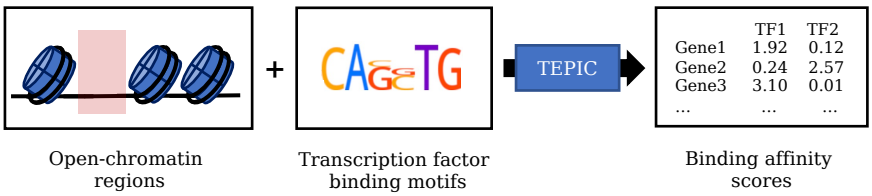


Figure 89: Overview of the TEPIC workflow.

8.1.3.2 *Scenario 5: Predicting gene expression values using* INVOKE

The inputs for Scenario 5 are paired open-chromatin regions and gene expression values. These can then be used to conduct the INVOKE (identification of key regulators) analysis of the TEPIC framework. This analysis consists of two steps. First, TEPIC is applied to the open-chromatin regions to calculate binding affinity scores for each regulator and target gene, as described in Section 8.1.3.1.
The resulting affinity scores are then used as features in a linear model with elastic net regularization to predict the provided gene expression values (cf. Section 3.8.2.2). To train the model, we first conduct a nested cross-validation procedure that evaluates the predictive performance. Here, the inner cross-validation is used to find the best the ratio between ridge and lasso penalty (cf. Section 3.8.2.2). The outer cross-validation is used to evaluate the performance of the model. For this purpose, we report three performance measures that are averaged across the folds of the outer cross-validation, i.e., Pearson's correlation coefficient [408], Spearman's rank correlation coefficient [504], and the mean squared error (MSE). All three measures can be used assess the validity of the results.
Additionally, our web service trains a new model on the entire data set, which is then used to produce the final results. Here, we consider all regulators with regression coefficient unequal to zero as influential regulators.

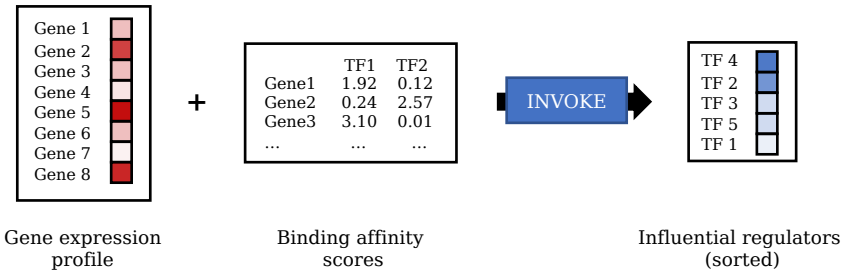


Figure 90: Overview of the INVOKE workflow.

8.1.4 *Analyzing key regulators in macrophages*

In this section, we evaluate the motif-based functionality of RegulatorTrail by analyzing a data set with paired chromatin accessibility (DNAse-seq) and gene expression (RNA-seq) profiles of macrophages from venous blood. The data set was obtained from the BLUEPRINT project [334] (Accession: S001S7).
Macrophages are phagocytes with many important roles in both the innate and adaptive immune system (cf. Section 2.2). In addition to their phagocytic activity, they are key players in antigen presentation [541] and even produce many essential cytokines [22, 86]. Due to their involvement in crucial immune processes, it is important to study regulatory mechanisms in these leukocytes.
In order to detect key regulators in the investigated macrophage sample, we uploaded both open chromatin regions and gene expression data to our web service and then conducted the following processing steps (Scenario 5). First, we used TEPIC to calculate affinity scores for each gene and all regulators based on the chromatin accessibility data and our complete set of transcription factor motifs (cf. Section 3.2.4.2). For the predicted affinity scores, we then trained a linear model with elastic net penalty to predict the provided gene expression values and to find the most influential regulators.
All parameters of the analysis are described in Section E.7. By comparing predicted and measured expression values, the best model in the outer cross-validation achieved a mean-squared error (MSE) of 0.623, a Pearson correlation of 0.616, and a Spearman correlation of 0.666.

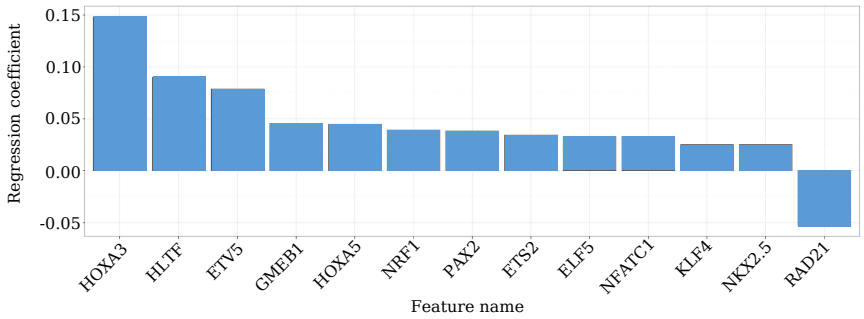


Figure 91: Barplot illustrating the INVOKE results for the analyzed macrophage sample. Each bar represents the regression coefficient for the respective regulator in the elastic net model. This Figure was adapted from [533].

The INVOKE analysis resulted in 13 transcription factors with a regression coefficient of at least 0.025 (cf. Figure 91). All 13 have already been discussed in the context of macrophages. Links to the respective publications are shown in Supplementary Table 48. In the following, we briefly discuss some of the key results.

The regulator with the highest score in our analysis is HOXA3. This transcription factor is known to promote maturation of macrophages [13]. Additionally, it has been shown that deregulated development in human macrophages of diabetic patients can be repaired by transduction of HOXA3 [17]. A further important transcription factor is ETS2, which is known to suppress inflammatory cytokines [323] and has been shown to promote metastasis in tumor-associated macrophages [599]. Another key regulator is the zinc finger protein KLF4, which seems to have essential functions in differentiation [480] and polarization [304] of macrophages.

8.2 DISCUSSION AND CONCLUSION

In this chapter, we presented the RegulatorTrail web service, a comprehensive tool suite for the evaluation and assessment of transcription factors, co-factors, and chromatin modifying proteins. In contrast to other tools, which are often created for one specific task, we designed our web service as a modular framework with various methods that can be applied to a broad range of application scenarios and input data types. A comparison of RegulatorTrail and other related web services is shown in Figure 92. As depicted, our web service provides not only more functionality, but also supports more databases and organisms than any of the other web services.

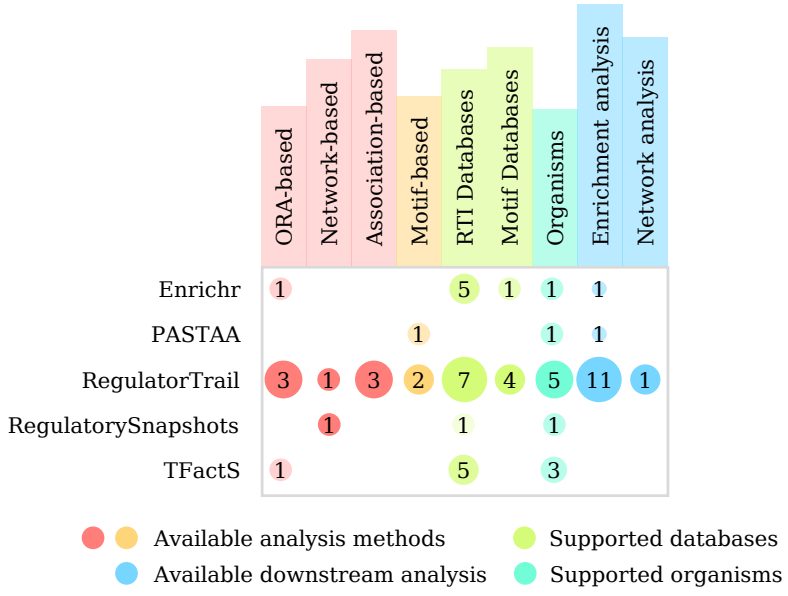


Figure 92: Comparison between RegulatorTrail and other approaches based on the number of available methods, supported databases, organisms, and downstream analyses.

The modularity of our framework allows us to easily extend the web service with additional methods and new features, such as the network-based approach (Scenario 3) that was not part of the original web service. Furthermore, we developed RegulatorTrail as part of the Graviton framework (cf. Chapter 4), which allows for a seamless integration with its sister projects GeneTrail and NetworkTrail that can directly be used in downstream analyses.

While RegulatorTrail already is one of the most powerful tools for the analysis of transcriptional regulators, there are still elements that can be improved or extended. In particular, our framework currently does not cover the full spectrum of possible application scenarios, like the analysis of genetic variations that effect binding patterns of transcription factors [388, 535, 612]. Additionally, neither RTI-based nor motif-based methods currently support information from distal enhancer regions. Instead regulators are assigned to a target gene if it has a binding site in a window around the TSS, which includes the promoter and proximal enhancers. While regulatory elements often strongly affect the nearest gene [139], it is still a simplified approach that could potentially be improved, e.g., by incorporating additional databases, like GeneHancer [154], or by using further data sources, such as Hi-C [543].

Nevertheless, the rich functionality of our framework combined with the flexibility and user-friendly web interface make RegulatorTrail one of the most comprehensive tools for the detection of influential transcription factors, chromatin modifying proteins, and associated biological mechanisms and set it apart from other approaches.

9

TCF3 AS MASTER REGULATOR IN BLASTEMAL WILMS TUMORS

Author contributions

This chapter is based on our publication "The role of TCF3 as potential master regulator in blastemal Wilms tumors" [532]. The wet lab experiments were conducted by Kathrin Katter, Nicole Ludwig, and Jenny Wegert. The computational analysis described in this chapter and the interpretation of the corresponding results has mainly been conducted by Hans-Peter Lenhof and me with additional contributions by Norbert Graph and Eckart Meese. The complete list of contributors can be found in the author list of the manuscript [532].

Wilms' tumors (WTs) are the most abundant type of kidney tumors in children [111]. While they generally have a good prognosis, around 13% of affected patients show a relapse within two years [166].
Over the years, two distinct treatment schemes have been developed [124]. The Children's Oncology Group (COG) prefers a nephrectomy followed by potential chemotherapy. In contrast, the International Society of Pediatric Oncology (SIOP) advocates for neoadjuvant chemotherapy before the actual surgery to reduce the tumor size. This facilitates the surgical removal of the tumor and also reduces the risk of tumor spillage during the procedure [124, 188].

After the nephrectomy, both SIOP and COG assess a variety prognostic markers to stratify patients concerning their risk of relapse [90]. The patients then receive a therapy that is adapted according to this risk assessment. This is a crucial step, because it has been shown that an intensive treatment regimen after surgery can have severe late effects [90]. In this context, it has been observed that an reduction of therapy for most patient subgroups can lead to more and healthier survivors [90]. Consequently, both protocols include clinical and molecular markers into risk assessment after surgery, which allows for a "risk-directed therapy" [124]. An overview of the different prognostic markers is shown in Figure 93.
In the following, we focus on prognostic markers that are relevant for cancers that are treated according to the SIOP protocol. Here, the preoperative chemotherapy is known to impact the histology patterns of Wilms' tumors, which generally consists of blastemal, epithelial, and stromal cells [570]. The composition of residual cells after neoadjuvant chemotherapy seems to be associated with the malignancy of the disease [570]. In this context, especially tumors that mainly persist of blastemal cells after chemotherapy seem to have a more adverse prog-

nosis. Consequently, the absolute volume of residual blastemal cells is discussed as a putative prognostic marker in future risk stratification schemes [124]. In order to further our understanding of blastema as a putative prognostic marker, it is crucial to study cellular processes that differentiate blastemal from non-blastemal components of tumors after the preoperative chemotherapy. For simplicity reasons, we hereafter refer to these tumors as either blastemal WTs or the blastemal subtype.

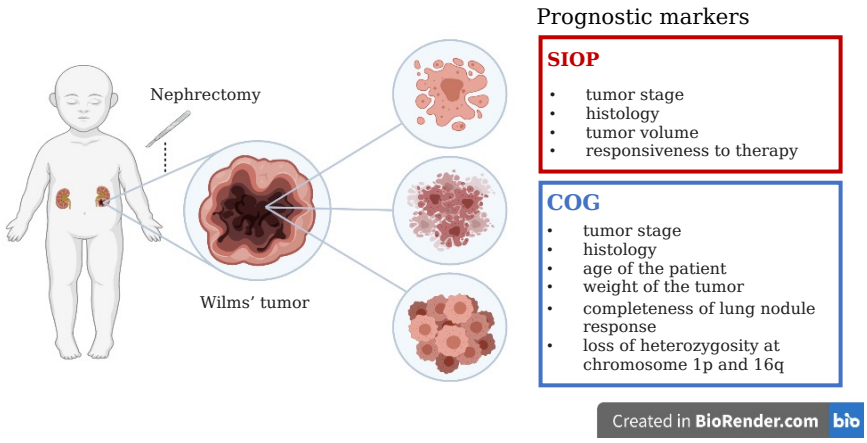


Figure 93: Overview of prognostic markers used for risk stratification of Wilms' tumors. The prognostic markers were obtained from [124] and the figure was created using BioRender.com.

In this context, several studies investigated molecular factors that contribute to the elevated malignancy of WTs with blastemal subtype. For example, although WTs rarely exhibit genetic alterations, Wegert et al. identified several mutations that occurred in around 18% of profiled blastemal WTs. Among them are mutations in transcription factors involved in developmental processes SIX1, SIX2, and MYCN, as well as miRNA-processing genes DGCR8 and DROSHA [568]. Additionally, it has been shown that the expression patterns of various miRNAs are altered in blastemal WTs compared to normal tissue [319].

Moreover, Wilms' tumors have also been used to study differences between tumor cells and healthy cells in the fetal kidney developmental process [296]. The corresponding results show that the blastemal component of WTs has expression patterns similar to cells in an early stage of metanephric kidney development, which suggests that the differentiation process in the corresponding cells might be impeded [296]. This observation is further supported by studies that analyzed WTs to investigate the roles of different stem cell factors during kidney development [114, 356]. Furthermore, using human xenografts of WTs, a subpopulation of blastemal cells has been observed that exhibit stem cell properties [414, 486].

The results of the previously described studies already indicate that

several (epi)genetic regulators and different stem cell factors might contribute to the elevated malignancy of blastemal subtype WTs. However, a systematic evaluation of transcriptional regulators that might drive the pathogenesis of the respective WT subtype had not yet been conducted. For this purpose, we analyzed gene expression profiles of 33 WT biopsies (17 blastemal and 16 non-blastemal) after neoadjuvant chemotherapy to study regulatory mechanisms that are characteristic for the blastemal subtype and that might be involved in the aggressiveness and resistance to chemotherapy of these tumors (cf. Section 9.1). Additionally, since many of our results indicate an essential role of transcriptional regulators involved in chromatin signaling, we also compared histone modification patterns of two cell cultures, one originating from blastemal and one from stromal subtype WTs, to study differences in their chromatin structure (cf. Section 9.2).

9.1 ANALYSIS OF GENE EXPRESSION PROFILES

For the gene expression analysis, we collected 33 Wilms tumor biopsies from patients that underwent neoadjuvant chemotherapy (SIOP treatment regimen). In total, the data set contains microarray profiles of 17 blastemal and 16 non-blastemal tumors. Both groups are matched in age (57 vs. 61.4 months) and gender (70% vs. 69% females). All processing steps, parameters, clinical details of patients, and ethics statements can be found in Appendix E.8.

9.1.1 *Influential regulators in blastemal Wilms' tumors*

First, we analyzed transcriptional regulators that might contribute to the increased malignancy of Wilms' tumors with predominantly blastemal subtype. To this end, we first used the Shrinkage t-test (cf. Section 3.4.3.2) to calculate expression differences between the two groups. We then created ten test sets from the resulting score list, i.e., all significantly up- and downregulated genes ($P < 0.01$), and the 250, 500, 750, and 1,000 genes with highest and lowest t-scores, respectively. For all gene lists, we then carried out a REGGAE analysis (cf. Chapter 7) to identify the most influential transcriptional regulators. We aggregated the result lists of the five upregulated and the five downregulated gene lists using the second-order statistic [123]. The aggregated p-values were then FDR-adjusted using the method proposed by Benjamini and Yekutieli (cf. Section 3.3.2).
In total, we analyzed 1,076, transcriptional regulators. Of these, 138 regulators seem to have a significant influence on upregulated genes and 112 on downregulated genes. The top 50 candidates of both analyses are shown in Table 7 and 8.

(A) Regulator	P-value	(B) Regulator	P-value
RUNX1 (-)	1.22×10^{-180}	NR2F2 (-)	7.83×10^{-116}
TCF3 (+)	5.96×10^{-163}	MAX (+)	3.27×10^{-105}
NR2F2 (+)	6.19×10^{-163}	TCF3 (-)	3.12×10^{-95}
MAX (-)	3.54×10^{-157}	RUNX1 (+)	1.78×10^{-94}
SFPQ (+)	1.06×10^{-136}	CREBBP (-)	8.51×10^{-78}
ELF1 (-)	4.60×10^{-134}	ELF1 (+)	1.09×10^{-76}
KDM5B (+)	1.68×10^{-131}	SUMO2 (-)	4.03×10^{-74}
HDAC1 (+)	9.85×10^{-125}	CREB1 (-)	4.42×10^{-70}
SIN3A (+)	2.90×10^{-123}	SMC3 (-)	8.33×10^{-70}
CREB1 (+)	5.84×10^{-123}	UBTF (-)	9.24×10^{-61}
SMC3 (+)	9.37×10^{-120}	RAD21 (-)	4.72×10^{-65}
CREBBP (+)	7.36×10^{-119}	HDAC1 (-)	1.03×10^{-61}
SUMO2 (+)	5.37×10^{-115}	SMARCC2 (-)	3.84×10^{-61}
RAD21 (+)	7.93×10^{-113}	SFPQ (-)	5.60×10^{-60}
FOXP1 (-)	3.66×10^{-104}	FOXP1 (+)	8.87×10^{-59}
STAT1 (-)	8.03×10^{-104}	KDM5B (-)	5.10×10^{-56}
UBTF (+)	5.37×10^{-102}	STAT1 (+)	2.86×10^{-55}
ZNF384 (+)	2.97×10^{-101}	SMAD3 (-)	1.13×10^{-50}
SMARCC2 (+)	1.08×10^{-94}	SIN3A (-)	1.99×10^{-50}
ERG (-)	2.55×10^{-90}	TAF7 (-)	1.78×10^{-49}
TAF7 (+)	6.93×10^{-90}	ZNF384 (-)	2.22×10^{-49}
SPI1 (-)	4.17×10^{-84}	SPI1 (+)	6.10×10^{-47}
HDAC2 (+)	4.41×10^{-82}	CEBPB (+)	1.22×10^{-46}
SMAD3 (+)	1.26×10^{-79}	ERG (+)	2.65×10^{-38}
HOXA4 (+)	2.65×10^{-75}	RUNX3 (+)	9.67×10^{-37}
SIX5 (+)	3.44×10^{-75}	CTCF (-)	3.51×10^{-35}
CEBPB (-)	2.90×10^{-70}	SIX5 (-)	2.60×10^{-33}
WDR5 (+)	3.43×10^{-66}	HOXA4 (+)	5.38×10^{-33}
KDM4A (+)	1.07×10^{-64}	FOSL1 (+)	9.67×10^{-37}
BMI1 (+)	1.94×10^{-64}	STAT5A (+)	3.04×10^{-31}
SP4 (+)	2.06×10^{-60}	GABPA (-)	4.81×10^{-31}
YY1 (+)	3.37×10^{-60}	BATF (+)	8.99×10^{-31}
BATF (-)	1.39×10^{-54}	HDAC2 (-)	2.48×10^{-30}
CTCF (+)	2.68×10^{-53}	YY1 (-)	5.81×10^{-30}
RUNX3 (-)	3.57×10^{-51}	VDR (+)	1.56×10^{-29}
STAT5A (-)	7.98×10^{-51}	NR2F1 (-)	2.12×10^{-29}
HOXA6 (+)	1.82×10^{-50}	NFATC1 (+)	1.25×10^{-28}
MTA3 (+)	8.70×10^{-49}	IKZF1 (+)	7.73×10^{-28}
GABPA (+)	4.30×10^{-46}	FOSL2 (+)	5.29×10^{-26}
CTBP2 (+)	8.20×10^{-46}	CTBP2 (-)	1.48×10^{-24}
SMARCC1 (+)	3.08×10^{-45}	PPARD (+)	3.03×10^{-24}
FOSL2 (-)	2.21×10^{-44}	KDM4A (-)	4.07×10^{-24}
KLF1 (-)	6.81×10^{-44}	MTA3 (-)	9.88×10^{-24}
NFATC1 (-)	8.68×10^{-44}	EP300 (-)	1.23×10^{-23}
MAFK (-)	2.56×10^{-43}	HOXA6 (-)	2.32×10^{-23}

Table 7: The top 50 transcriptional regulators obtained by REGGAE (Rank 1 to 45) for (A) the most upregulated and (B) the most downregulated genes (blastemal vs. non-blastemal WTs). Both colors of regulators and the sign in parentheses indicate if the regulator has a **positive** or **negative** correlation with its target genes. This table was adapted from [532].

9.1.1.1 *The top regulators*

Amongst the top four regulators in both analyses, we identified NR2F2, TCF3, RUNX1, and MAX. NR2F2 and TCF3 have a positive correlation with upregulated genes in blastemal WTs and RUNX1 and MAX a negative one. In the following paragraphs, we briefly describe their general function and, in particular, if they have already been discussed in the context of stem cell development, cancer initiation, or cancer progression.

NR2F2 (or COUP – TF2) is a family member of the "steroid thyroid hormone superfamily of nuclear receptors" [379]. Amongst others, this transcription factor is known to be involved in the regulation of embryonic stem cell (ESC) differentiation [449]. Furthermore, it has been shown that this nuclear receptor can directly promote cancer hallmarks, such as angiogenesis (cf. Section 2.3.1.1) or metastasis (cf. Section 2.3.1.1) [424].

TCF3 is part of the TCF/LEF family of transcription factors. These regulators are involved in the Wnt signaling pathway, which plays important roles in embryogenesis [19] and cancer development [605]. Moreover, the activation of this pathway is regularly observed in Wilms' tumors [102] and, in particular, in the blastemal subtype [510]. Additionally, TCF3 has been discussed as one of the central regulators in embryonic stem cells [97]. Besides TCF3 itself, our results list also contains CREBBP and EP300, which are known coactivators of TCF3 [82].

RUNX1 (or AML1) is a transcription factor that controls the development and maintenance of blood cells [232] as well as their differentiation into myeloid or lymphoid cells. [391]. Mutations of this regulator are associated with the aggressiveness of multiple cancer types [353, 429]. Moreover, RUNX1 has been described as tumor suppressor gene in acute lymphoblastic leukemia [115] and breast cancer [559].

MAX is a member of the "basic helix-loop-helix leucine zipper (bHLHZ) family of transcription factors" [378]. Generally, it forms homodimers or heterodimers with other members of the bHLHZ family, e.g., MYC, MXI, and MNT, which can associate with both coactivator and corepressor complexes to control a diverse set of cellular functions, including cell proliferation, differentiation, and apoptosis [189]. MAX itself has been discussed as a tumor suppressor gene in different cancer types [277, 328].

(A) Regulator	P-value	(B) Regulator	P-value
VDR (-)	6.54×10^{-43}	KLF1 (+)	3.00×10^{-23}
EP300 (+)	7.79×10^{-43}	SP4 (-)	6.82×10^{-23}
ZBTB33 (+)	2.59×10^{-39}	MAFK (+)	1.17×10^{-22}
NR2F1 (+)	5.83×10^{-39}	WDR5 (-)	1.80×10^{-22}
DUX4 (-)	4.33×10^{-37}	IRF4 (+)	1.38×10^{-21}

Table 8: The top 50 transcriptional regulators obtained by REGGAE (Rank 46 to 50) for (A) the most upregulated and (B) the most downregulated genes (blastemal vs. non-blastemal WTs). Both colors of regulators and the sign in parentheses indicate if the regulator has a **positive** or **negative** correlation with its target genes. This table was adapted from [532].

9.1.1.2 *Regulators involved in cancer development and progression*

Amongst the top candidates in our REGGAE results, we also identified several other transcriptional regulators that are directly associated with oncogenesis in different cancer types and, hence, might help to explain the elevated malignancy of blastemal WTs. These include regulators with known roles in tumor initiation (e.g., ERG5 [113] and FOXP1 [199]), transcription factors associated with tumor progression (e.g., ELF1 [565] and STAT1 [263]), proteins that may contribute to the development of metastases (e.g., CREB1 [560] and KDM5B [521]), or the resistance to chemotherapy (e.g., SFPQ [406] and BMI1 [488]).

Furthermore, we identified several regulators that promote epithelial-to-mesenchymal transition (EMT) in different cancer types, such as KDM5B [521], or SMAD3 [590]. EMT is a process during normal embryogenesis that is essential for tissue development [250, 445]. It allows epithelial cells to transform into a mesenchymal phenotype [250]. Mesenchymal cells are multipotent cells that can differentiate into various cell types [250]. Hence, respective cells exhibit an increased resistance to apoptosis, reduced cell adhesion, and gain the ability to migrate [250, 445]. Accordingly, this process is also regularly be found to be active in cancer cells [250, 445].

Moreover, we also find many transcriptional regulators that are involved in chromatin signaling or remodeling, such as SIN3A [193], HDAC1, and HDAC2 [455]. Chromatin remodeling complexes are essential regulatory components in embryonic stem cells that mediate between differentiation and pluripotency [255]. Disruptions in chromatin signaling are regularly linked to oncogenesis [557, 558].

9.1.2 *Influential regulator complexes*

Generally, transcriptional regulators do not work alone. Instead, they cooperate with other regulators or cofactors to form larger protein complexes to control the expression of their target genes. In order to incorporate complex information in our analysis, we examined if regulators that belong to certain protein complexes are significantly enriched in the sorted REGGAE result lists. To this end, we applied gene set enrichment analyses for biological categories extracted from protein complex databases (i.e., CORUM [454] and EpiFactors [350]) as well as custom gene sets derived from literature.
Since many of the identified regulators in our REGGAE analysis are involved in chromatin signaling, developmental processes, the regulation of stem cells, or stem cell properties in cancer cells, we focused on gene sets that represent regulatory functions in ESCs, i.e., chromatin signaling [248], pluripotency states [526], and the occupation of known super-enhancers [212]. All processing steps and the complete set of parameters can be found in Appendix E.8.5.

9.1.2.1 *Regulator complexes involved in chromatin signaling*

Indeed, the GSEA results confirm our observations that chromatin-modifying proteins could play an important role in blastemal subtype WT. We identified several enriched categories associated with chromatin remodeling complexes. Amongst others, these include the ATP-dependent chromatin remodeling complexes, such members of the SWI/SNF family or the NuRD complex. SWI/SNF (switch/sucrose non-fermentable) complexes are involved in the repositioning, ejecting, and incorporation of nucleosomes [94]. The NuRD (nucleosome remodeling and deacetylation complex) complex couples the nucleosome remodelling activity of the SWI/SNF proteins CHD3 and CHD4 with the histone deacetylases HDAC1 and HDAC2 [119].
Our observations are further supported by our analysis of the chromatin signaling network discussed by Juan et al. [248]. This network contains regulators that control essential functions in the development and maintenance of ESC, such as chromatin remodeling, differentiation, pluripotency, or stemness [248]. The corresponding gene set is highly enriched in both REGGAE results lists with a p-value of 7.34×10^{-4} for down-regulated genes and 8.49×10^{-5} for upregulated ones. Accordingly, it contains many top candidates of our analysis, in particular, TCF3 seems to be one of the central regulators in this context.

In order to further elucidate the impact of TCF3 in this communication network, we collected binding sites of this regulator from our RTI database and Chip-seq experiments in mouse ESCs [97, 333]. These were then used to study putative regulatory interactions be-

tween TCF3 and other genes in this network presented by Juan et al [248]. Interestingly, of the 49 genes, all seem to have respective binding sites, and 35 additionally have a high correlation with the expression of TCF3 in our microarray profiles ($|\rho| > 0.5$). The respective genes are depicted in Figure 94. In summary, these results provide further evidence that chromatin signaling plays an important role in blastemal WTs and that TCF3 might be one of the key regulators in this context.

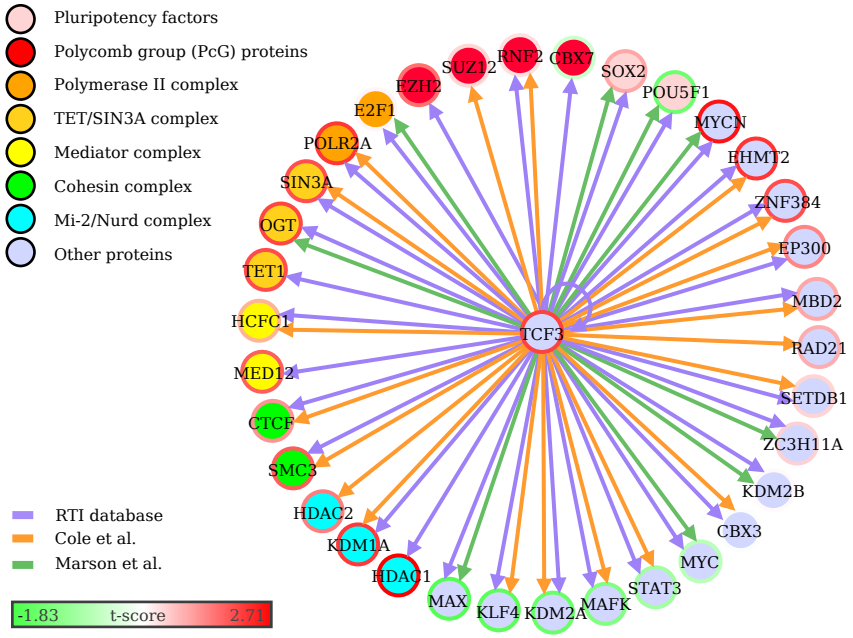


Figure 94: Putative TCF3 target genes in the chromatin signaling network of ESCs [248]. Each vertex represents putative TCF3 targets in this network, and edges indicate if these genes have TCF3 binding sites and a strong absolute correlation ($|\rho| > 0.5$) with the regulator. The color of each gene represents its assignment to certain protein complex, and the color of its border indicates the corresponding t-score (blastemal vs. non-blastemal WT). The edge color depicts the source of respective binding site. This figure was adapted from [532]

9.1.2.2 *Further relevant regulator complexes*

Besides the many regulators involved in chromatin signaling, our results also feature additional significantly enriched functional categories. This includes complexes that are involved in the regulation of developmental processes in ESCs, such as SIN3 [193] and SMAD [325]. Additionally, we analyzed proteins that mediate pluripotency states in ESCs. This analysis revealed that regulators that promote a poised state of developmental genes are significantly enriched. Furthermore, our results also highlight that many of the top-ranked

regulators are known to occupy super-enhancers in embryonic stem cells, like EP300, CREBBP, CTCF, HDAC1, HDAC2, RAD21, SMC3, SMAD3, STAT3, and TCF3.

9.1.3 *Analysis of kidney developmental genes*

The results of previous studies show that the triphasic histology patterns of Wilms' tumors is shared by different cells in early stages of kidney development [442]. Since our analysis of the most influential regulators also highlighted the role of different components associated with the control of developmental processes, we conducted further experiments in this direction.
For this purpose, we analyzed if biological categories directly involved in kidney development or associated biological processes are significantly enriched amongst the 100 top-ranked regulators in our REGGAE analysis and the 1,000 most upregulated genes in blastemal WTs. To this end, we conducted over-representation analysis for custom gene sets extracted from literature [78, 512], and categories from the GeneTrail collection, i.e., GO [100], KEGG [394], and WikiPathways [256]. All processing steps and the complete set of parameters can be found in Appendix E.8.5.
Our enrichment analysis for the most upregulated genes in blastemal WTs identified a significant enrichment of genes involved in the development, differentiation, and morphogenesis of kidney cells. In this context, we especially found an enrichment of genes that are active in specific components of embryonic kidneys, i.e., the cap mesenchyme or the metanephric mesenchyme [78]. Accordingly, we identified various enriched biological categories that are associated with different stages of kidney development and morphogenesis. These observations confirm results obtained by Li et al., suggesting that retained blastemal cells after neoadjuvant chemotherapy could potentially resemble cells in an early stage of the metanephric mesenchymal-epithelial transition [296]. Additionally, we also detected an enrichment of the non-canonical Wnt pathway, which is an important signaling pathway in blastemal WTs and WTs in general [165, 510].
In our analysis of the most influential regulators, we identified a significant enrichment of the TGF-beta receptor signaling pathway, which inter alia is involved in cell growth and differentiation. In accordance with the results for the most upregulated genes, we also identified an enrichment of the Wnt signaling pathway and especially sub-components that control pluripotency and self-renewal in ESCs.
In summary, both analyses support previous observations suggesting that cells of the blastemal subtype seem to resemble cells during early kidney developmental stages and could potentially have stem cell properties.

9.2 ANALYSIS OF HISTONE MARKS

Many of our results indicate an important role of chromatin remodeling complexes and stem cell factors in blastemal cells that survive the chemotherapy. This is in line with previous studies that observed subpopulations of cells in blastemal WTs that might have cancer stem cell characteristics [414, 486]. These results suggest that the chromatin structure in the blastemal subtype could be altered compared to other subtypes. In order to further elucidate this assumption, we measured histone modification patterns, i.e., H3K4me3 and H3K27ac marks, in two cell cultures that were created from primary WT cells. One cell culture exhibited mesenchymal characteristics and was derived from a tumor with predominantly stromal histology (*ws*568li) [567], and the second cell culture has blastemal characteristics and was derived from a mouse xenograft of a WT with triphasic histology pattern (*ws*998M18). The processing steps for the respective ChIP-seq experiments and all parameters can be found in Appendix E.9. For comparison, we also use ChIP-seq experiments of human ESCs from the Roadmap Epigenomics Mapping Consortium (Epigenome E015) [52].

9.2.1 *Comparison of histone marks in human ESCs and WT cells*

In order to compare the histone modification patterns across the three cell types, we analyzed if the promoter regions of genes that carry the respective epigenetic marks, i.e., H3K4me3 and H3K27ac. To this end, we define promoter regions as a window around the transcription start site (TSS) of a gene (TSS +/- 5kb). For each promoter region, we checked if it overlaps with both analyzed histone marks. All processing steps and parameters of this analysis are described in Appendix E.9. An overview of the results is depicted in Figure 95.
First, we analyzed the H3K4 trimethylation patterns across all protein-coding genes in our data set. Here, we detected 19,532 genes that carried this histone mark in all three cell lines (cf. Figure 95A). Of these, 11,886 also contain H3K27 acetylations in all samples, while 3,845 are only available in ESCs and blastemal cells (cf. Figure 95B). For the latter group, we conducted an over-representation analysis to identify significantly enriched biological processes. The respective results show that the genes seem to be associated with proliferation or important signaling pathways, most notably the Wnt-signaling pathway. Additionally, we observe an enrichment of categories with marker genes of specific cancer types, e.g., gastric cancer, glioblastoma, retinoblastoma.

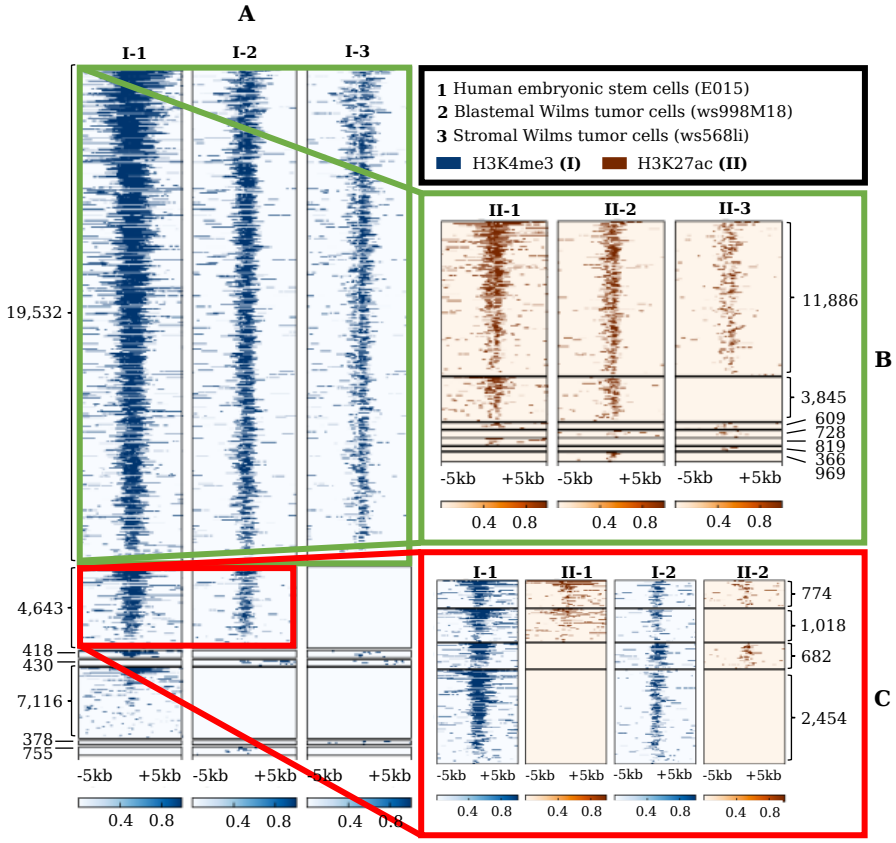


Figure 95: Comparison of histone marks in the promoter regions of (1) human embryonic stem cells (ESC), (2) blastemal WT cells, and (3) mesenchymal WT cells. (A) Heatmap depicting H3K4me3 patterns across known promoter regions of all genes annotated in the GENCODE database [207]. (B) Heatmap illustrating H3K27ac patterns in the promoter region of genes with H3K4me3 marks across all samples. (C) Heatmap with combined H3K4me3 and H3K27ac marks in ESCs and blastemal WTs that are not present in mesenchymal cells. This figure was adapted from [532].

Additionally, we identified 4,643 genes that carry H3K4me3 modifications in ESCs and blastemal cells, but not in mesenchymal cells. As shown in Figure 95C most of these regions do not contain a H3K27ac mark, which indicates that these promoter regions could be in a putative poised chromatin state. An over-representation analysis of affected genes showed an enrichment of cell fate, cell specification, development, and differentiation processes.

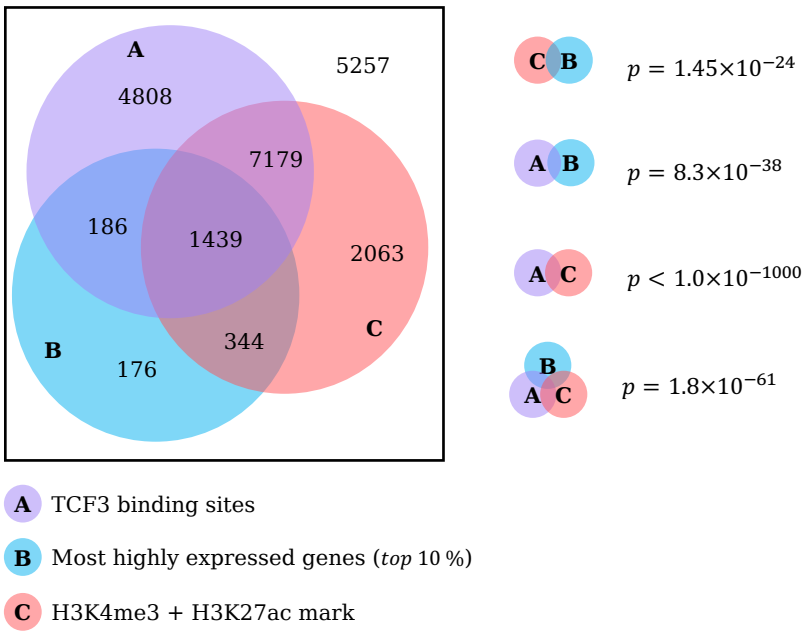


Figure 96: Venn diagrams depicting the overlaps between (1) TCF3 binding sites, (2) genes with active chromatin state, and (3) the most upregulated genes in blastemal WTs. This figure was adapted from [532]

9.2.2 *Integrative analysis of epigenomic and transcriptomic data*

Many of the results presented in the previous sections indicate that WTs with blastemal subtype could have retained stem cell properties. We even identified TCF3 as one of the key regulatory elements in this context. In order to further support these assumptions, we analyzed if there are significant overlaps between (1) the most upregulated genes in blastemal WTs (top 10%), (2) genes with an active chromatin state, and (3) genes with TCF3 binding sites. For this purpose, we conducted ChIP-seq experiments for TCF3 in our cell line with blastemal characteristics.
Although we repeated this experiment with different antibodies, the respective results were extremely noisy and could unfortunately not be utilized to examine TCF3 binding sites. Instead, we decided that the best alternatives are TCF3 ChIP-seq experiments in mouse ESCs [97, 333], which we then mapped to human orthologs.
To study the overlaps between the three gene sets, we applied a hypergeometric test. For this purpose, we used all genes measured in our gene expression analysis as reference. As depicted in Figure 96, all pairwise comparisons, as well as the intersection of all three data, types provided highly significant overlaps.

9.3 DISCUSSION AND CONCLUSION

While the majority of children diagnosed with nephroblastoma have a good prognosis, approximately 13% exhibit a relapse within two years. One important risk factor that indicates an unfavorable prognosis are blastemal cells that resist the preoperative chemotherapy. For this reason, it is crucial to understand the role of these cells and, in particular, to identify potential biomarkers that could improve diagnosis, risk assessment, and therapy of Wilms' tumors.

In this chapter, we described a study in which we compared blastemal tumors and other subtypes to identify influential regulators and associated mechanisms that might control the increased aggressiveness and resistance to chemotherapy of blastemal WTs [532].

In our analyses, we identified many influential regulators involved in processes like chromatin remodeling, embryonic development, and the orchestration of pluripotency. Here, especially, regulators that are contained in the chromatin signaling network of ESCs described by Juan et al. [248] seem to play an important role. In particular, our analyses provide strong evidence that the transcription factor TCF3 is one of the central regulators in this context. TCF3 is also a key regulator of the Wnt signaling pathway, which has been identified as a crucial process in blastemal WTs [165, 338, 510]. Here, it interacts with central pluripotency factors POU5F1, NANOG, and SOX2 to control cell growth and self-renewal [493]. This is also confirmed by our ORA results that highlight significantly enriched categories associated with the Wnt pathway and pluripotency.

Additionally, our results confirm observations of previous studies indicating that blastemal WTs resemble cells in early stages of the kidney development [114, 296, 356] or might even exhibit stem cell properties [414, 486]. Although previous results already highlighted that blastemal WTs exhibit stem cell-like properties when compared to normal kidney cells [11], we compared histone modification patterns of two cell lines derived from primary tumors and embryonic stem cells to reinforce these observations. These analyses revealed that blastemal cells indeed share characteristics of stem cells that are not present in other subtypes, which could potentially contribute to their elevated malignancy. A comparison of genes with activating histone marks in their promoter regions, the most upregulated genes in blastemal WT, and target genes of TCF3 further confirmed the role of TCF3 as a crucial regulator in these cells.

In conclusion, our study confirms and reinforces previous observations about blastemal cells that survive neoadjuvant chemotherapy. The presented results clearly emphasize that developmental processes in blastemal WTs seem to be impeded and that respective cells exhibit essential stem cell characteristics that are not present in other WT subtypes. We also identified a circuitry of regulatory mechanisms that seem to control these properties and that could even be involved in the elevated malignancy and resistance to chemotherapy of blastemal WTs. As a central component in this context, we identified and discussed the crucial role of TCF3. Besides TCF3, our analysis revealed other putative marker genes of WTs with blastemal subtype. In the future, our insights could potentially lead to an improved risk assessment and therapy stratification for Wilms' tumors patients.

10 SUMMARY, DISCUSSION, AND CONCLUSION

In the last three decades, high-throughput techniques, such as next-generation sequencing, microarrays, or mass spectrometry, have revolutionized biomedical research by generating detailed molecular profiles of biological samples. Furthermore, they have become essential tools in non-research settings, such as forensics [77] or clinical applications [308].
The data sets created by the different high-throughput platforms are typically high-dimensional and noisy, making manual inspections impossible. Hence, powerful computational methods are required to analyze these complex data sets.
In this thesis, we presented a comprehensive framework of algorithms, tools, and databases that facilitate an integrative analysis of molecular high-throughput profiles. We developed these tools with the major goal to investigate biological processes that are deregulated in complex diseases like cancer and to identify potential driving factors within those processes.
In the following sections, we summarize and discuss the presented work and highlight possible directions for future research.

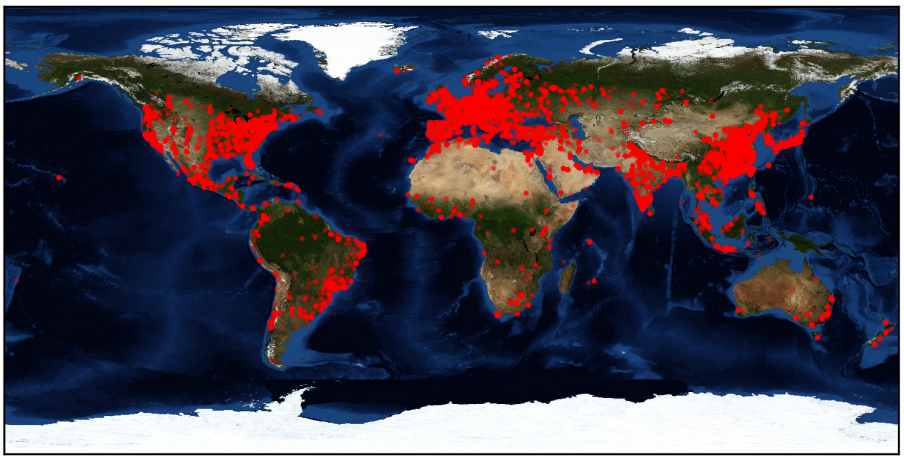


Figure 97: Combined geolocation data of users of all web services presented in this work. The data was collected between May 2016 and Januanary 2022.

10.1 SUMMARY

All tools and databases presented in this work were developed as web services. To make this possible, we created a web framework called Graviton that builds the basis of all our tools (cf. Chapter 4). By using a common framework, we reduced the overhead associated with the development and maintenance of different web applications. For example, we can easily share source code between our tools, which helps avoid unnecessary code duplication and makes updating and maintaining the individual tools much easier. For users, the usage of our web service ensures that all software components are compatible and that all data resources are properly curated and sanitized. Additionally, the usage of a web service removes or facilitates many technical challenges for users, such as the installation of third-party applications or the processing of external databases. Consequently, users can spend more time on the analysis of their data set. In the subsequent paragraphs, we shortly review and discuss the different algorithms, tools, and databases that were presented in this work. Additionally, we summarize the most important results of the conducted case studies.

First, we presented the GeneTrail web service, one of the most powerful enrichment and network analysis toolboxes available today (cf. Chapter 5). In comparison to other web services, GeneTrail excels in various aspects. First of all, our web service provides a powerful framework of enrichment and network analysis algorithms with highly efficient C++ implementations. The different approaches can be applied to explore our comprehensive collection of biological categories that incorporates 40 external data sources for 15 prevalent organisms. However, unlike most competing methods, GeneTrail cannot only be applied to analyze deregulated biological processes in traditional bulk data sets, but also to explore time-series experiments or single-cell data. To complement the rich functionality of our toolbox, we also created an intuitive web interface with many interactive visualizations. These range from a general overview of the computed enrichment results to an in-depth characterization of individual categories or signaling pathways.

In Chapter 5, we also presented the results of two studies that highlighted the capabilities of GeneTrail. First, we analyzed a single-cell RNA-seq data set of CD14 monocytes from the peripheral blood of COVID-19 patients and healthy controls [581]. Here, our results indicate that in patients with acute respiratory distress syndrome (ARDS), the activity of the adaptive immune system might be impeded, while processes of the innate immune system seem to be overactive. In a second study, we analyzed time-resolved gene expression profiles of CD4+ cells that were in vitro activated [121]. In particular, we investigated and discussed biological processes and associated genes with

distinct time courses after T cell activation. Both analyses successfully demonstrate that our web service can uncover relevant biological information from molecular high-throughput experiments that might help to gain novel insights into complex pathogenic mechanisms.
Using the enrichment analysis functionality of GeneTrail, we were also able to create miRPathDB, a web-based dictionary that compiles information about miRNAs, their target genes, and putative target pathways (cf. Chapter 6). Our database stands out by providing a tenfold increase of information compared to any competing resource. This was achieved by the integration of novel miRNAs and miRNA candidates from our miRCarta database [28]. Additionally, miRPathDB also offers different analysis tools that can be used to compare similarities between different miRNAs and to explore associations between miRNAs and target pathways.
Besides the identification of deregulated biological processes, we also focused on the detection of transcriptional regulators that might drive these deregulations. Here, our first contribution was REGGAE, a novel algorithm for the identification of key transcriptional regulators that have a significant effect on deregulated processes (cf. Chapter 7).
REGGAE uses an enrichment-based method to identify and rank regulators based on associations between regulators and differentially expressed target genes. We evaluated REGGAE and related algorithms in two different application scenarios. First, we compared expression profiles of estrogen receptor-positive (ER+) and estrogen receptor-negative (ER−) breast cancer cells to identify factors that might be responsible for gene expression differences between the two subtypes and that could contribute to the increased malignancy of ER− tumors. Secondly, we analyze perturbation signatures of specific regulators to examine if the considered methods could detect the perturbed regulators. The results of all conducted analyses demonstrate that REGGAE could outperform competing methods. In fact, our algorithm was the only approach that was able to uncover the most influential regulators in all cases. Furthermore, our approach provides supplemental information that facilitates the analysis and interpretation of the obtained results and sets REGGAE further apart from other approaches. This includes estimates whether the identified regulators might act as activators or repressors and different confidence measures, like P-values or confidence intervals, which help to assess the validity of the results.
For the analysis of influential transcriptional regulators, we also created the RegulatorTrail web service, one of the most comprehensive toolboxes for this purpose (cf. Chapter 8). In contrast to other tools that often only focus on one specific application, our web service offers various methods for identifying key regulators that can be applied in a broad range of application scenarios to analyze transcriptomic, proteomic, or epigenomic data sets.

In Chapter 9, we demonstrated the capabilities of our tools suite by conducting a detailed analysis of our Wilms' tumor (WT) data set. While WTs typically have a good prognosis, tumors that have a predominantly blastemal histology after neoadjuvant chemotherapy often seem to have a much more unfavorable outcome. In our study, we compared WTs with a blastemal subtype and tumors with different histology to identify factors that might drive the increased malignancy of blastemal tumors. In particular, our results confirm previous observations suggesting that blastemal WTs exhibit stem cell-like properties. Hence, we set out to find potential key players in this context. Indeed, we identified several putative biomarkers that are distinctive for the blastemal subtype. Many of them have essential roles in epigenetic processes that are already associated with increased malignancy in different cancer types. Among the top candidates, we identified TCF3 as one of the key regulators of these mechanisms. In the future, we hope that many of these insights can be utilized to improve therapy for Wilms' tumors.
Interestingly, our results were also discussed in a recent study by Zhou et al. that investigated if TCF3 could be a suitable therapeutic target for WTs [613]. To this end, the group silenced the regulator in G401 kidney tumor cells and assessed the viability of the tumor cells. Corresponding results showed that a knock-out of TCF3 reduces the activity of the Wnt pathway, significantly inhibits cell viability, reduces migration, and accelerates apoptosis in respective cells. Consequently, these results confirm our observations that TCF3 plays a crucial role in the malignancy of WTs and might even be a suitable candidate for future therapies.

10.2 PERSPECTIVES

The algorithms and tools discussed in this work already build a comprehensive framework for the analysis of molecular high-throughput profiles that is employed by many research groups around the globe (cf. Figure 97). However, several challenges remain that could be addressed in future research.
For example, while molecular profiling techniques, like high-throughput sequencing, have revolutionized biomedical research over the last two decades, their development is an ongoing process. New or updated protocols are emerging regularly, and computational approaches need to be adapted to support them. Here, one of the most promising developments is the increasing availability of single-cell multimodal omics protocols that make it possible to measure different omics types in the same cells [527]. While this poses new computational challenges, it also offers new possibilities, such as the ability to analyze biological processes using multiple measurement types si-

multaneously and to study interactions between them.
Moreover, our tool suite could be extended with workflows for additional omics types that are currently not supported. Promising candidates in this context would be long non-coding RNAs or metabolites. Both data types could provide additional insights into molecular mechanisms. Long non-coding RNAs can interact with DNA, RNA, or proteins to regulate gene expression [507]. Consequently, they have essential roles in many biological processes [92, 507]. Some of them have been associated with diseases like Alzheimer's, cancer, or diabetes [122, 144]. Furthermore, an inclusion of metabolic data could also provide additional information about the state of cells that is currently missing in our tool suite. These molecules constitute the substrate, products, or intermediates of many molecular reactions in the cell and, hence, an inclusion of metabolic measurements could further advance our understanding about cellular mechanisms [216, 311].
Although our web service are already highly interconnected such that users can seemlessly transition between different workflows, there are still places where the interoperability of our tools could be further expanded. For example, in the GeneTrail single-cell workflow, after a user has identified an interesting biological category, it might be helpful to investigate potential transcriptional regulators of the member genes using RegulatorTrail or even putative drug targets using DrugTargetInspector.

10.3 CONCLUSION

In conclusion, we presented a comprehensive tool suite with rich functionality for the analysis of molecular high-throughput profiles. The outlined methods can be applied to assess deregulated biological processes and driving factors across a wide range of application scenarios, which sets it apart from other approaches. Moreover, the presented case studies demonstrate that our framework can help elucidate complex pathogenic mechanisms and, hence, might be of broad interest to the scientific community.

Part I

APPENDIX

A

LIST OF PUBLICATIONS

A.1 FIRST AND JOINT FIRST AUTHOR PUBLICATIONS

Christina Backes, **Tim Kehl**, Daniel Stöckel, Tobias Fehlmann, Lara Schneider, Eckart Meese, Hans-Peter Lenhof, and Andreas Keller. "miRPathDB: a new dictionary on microRNAs and target pathways." In: *Nucleic acids research* (2016), gkw926.

Nico Gerstner, **Tim Kehl**, Kerstin Lenhof, Lea Eckhart, Lara Schneider, Daniel Stöckel, Christina Backes, Eckart Meese, Andreas Keller, and Hans-Peter Lenhof. "GeneTrail: a framework for the analysis of high-throughput profiles." In: *Frontiers in Molecular Biosciences* (2021), p. 890.

Nico Gerstner, **Tim Kehl**, Kerstin Lenhof, Anne Müller, Carolin Mayer, Lea Eckhart, Nadja Liddy Grammes, Caroline Diener, Martin Hart, Oliver Hahn, et al. "GeneTrail 3: advanced high-throughput enrichment analysis." In: *Nucleic Acids Research* (2020).

Tim Kehl, Christina Backes, Fabian Kern, Tobias Fehlmann, Nicole Ludwig, Eckart Meese, Hans-Peter Lenhof, and Andreas Keller. "About miRNAs, miRNA seeds, target genes and target pathways." In: *Oncotarget* 8.63 (2017), p. 107167.

Tim Kehl, Fabian Kern, Christina Backes, Tobias Fehlmann, Daniel Stöckel, Eckart Meese, Hans-Peter Lenhof, and Andreas Keller. "miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database." In: *Nucleic acids research* 48.D1 (2020), pp. D142–D147.

Tim Kehl, Lara Schneider, Kathrin Kattler, Daniel Stöckel, Jenny Wegert, Nico Gerstner, Nicole Ludwig, Ute Distler, Markus Schick, Ulrich Keller, et al. "REGGAE: a novel approach for the identification of key transcriptional regulators." In: *Bioinformatics* 1 (2018), p. 8.

Tim Kehl, Lara Schneider, Kathrin Kattler, Daniel Stöckel, Jenny Wegert, Nico Gerstner, Nicole Ludwig, Ute Distler, Stefan Tenzer, Manfred Gessler, et al. "The role of TCF3 as potential master regulator in blastemal Wilms tumors." In: *International journal of cancer* 144.6 (2019), pp. 1432–1443.

Tim Kehl, Lara Schneider, Florian Schmidt, Daniel Stöckel, Nico Gerstner, Christina Backes, Eckart Meese, Andreas Keller, Marcel H Schulz, and Hans-Peter Lenhof. "RegulatorTrail: a web service for the identification of key transcriptional regulators." In: *Nucleic acids research* 45.W1 (2017), W146–W153.

A.2 CO-AUTHOR PUBLICATIONS

Christina Backes, Tobias Fehlmann, Fabian Kern, **Tim Kehl**, Hans-Peter Lenhof, Eckart Meese, and Andreas Keller. "miRCarta: a central repository for collecting miRNA candidates." In: *Nucleic acids research* 46.D1 (2018), pp. D160–D167.

Caroline Diener, Martin Hart, **Tim Kehl**, Stefanie Rheinheimer, Nicole Ludwig, Lena Krammes, Sarah Pawusch, Kerstin Lenhof, Tanja Tänzer, David Schub, et al. "Quantitative and time-resolved miRNA pattern of early human T cell activation." In: *Nucleic Acids Research* (2020).

Michael Dirnberger, **Tim Kehl**, Tim Mehlhorn, Kurt Mehlhorn, and Adrian Neumann. "Towards an open online repository of P. polycephalum networks and their corresponding graph representations." In: *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS)*. 2016, pp. 588–588.

Michael Dirnberger, **Tim Kehl**, and Adrian Neumann. "NEFI: Network extraction from images." In: *Scientific reports* 5 (2015), p. 15669.

Tobias Fehlmann, Thomas Laufer, Christina Backes, Mustafa Kahramann, Julia Alles, Ulrike Fischer, Marie Minet, Nicole Ludwig, Fabian Kern, **Tim Kehl**, et al. "Large-scale validation of miRNAs by disease association, evolutionary conservation and pathway activity." In: *RNA biology* 16.1 (2019), pp. 93–103.

Martin Hart, Laura Nickl, Barbara Walch-Rueckheim, Lena Krammes, Stefanie Rheinheimer, Caroline Diener, Tanja Taenzer, **Tim Kehl**, Martina Sester, Hans-Peter Lenhof, et al. "Wrinkle in the plan: miR-34a-5p impacts chemokine signaling by modulating CXCL10 / CXCL11 / CXCR3-axis in CD4+, CD8+ T cells, and M1 macrophages." In: *Journal for Immunotherapy of Cancer* 8.2 (2020).

Martin Hart, Barbara Walch-Rückheim, Lena Krammes, **Tim Kehl**, Stefanie Rheinheimer, Tanja Tänzer, Birgit Glombitza, Martina Sester, Hans-Peter Lenhof, Andreas Keller, et al. "miR-34a as hub of T cell regulation networks." In: *Journal for immunotherapy of cancer* 7.1 (2019), p. 187.

Fabian Kern, Ernesto Aparicio-Puerta, Yongping Li, Tobias Fehlmann, **Tim Kehl**, Viktoria Wagner, Kamalika Ray, Nicole Ludwig, Hans-Peter Lenhof, Eckart Meese, et al. "miRTargetLink 2.0—interactive miRNA target gene and target pathway networks." In: *Nucleic Acids Research* (2021).

Fabian Kern, Lena Krammes, Karin Danz, Caroline Diener, **Tim Kehl**, Oliver Küchler, Tobias Fehlmann, Mustafa Kahraman, Stefanie Rheinheimer, Ernesto Aparicio-Puerta, et al. "Validation of human microRNA target pathways enables evaluation of target prediction tools." In: *Nucleic Acids Research* (2020).

Kerstin Lenhof, Lea Eckhart, Nico Gerstner, **Tim Kehl**, and Hans-Peter Lenhof. "Simultaneous regression and classification for drug sensitivity prediction using an advanced random forest method." In: *Scientific Reports* (2022).

Kerstin Lenhof, Nico Gerstner, **Tim Kehl**, Lea Eckhart, Lara Schneider, and Hans-Peter Lenhof. "MERIDA: a novel Boolean logic based integer linear program for personalized cancer therapy." In: *Bioinformatics* (2021).

Lara Schneider, Daniel Stöckel, **Tim Kehl**, Andreas Gerasch, Nicole Ludwig, Petra Leidinger, Hanno Huwer, Stefan Tenzer, Oliver Kohlbacher, Andreas Hildebrandt, et al. "DrugTargetInspector: An assistance tool for patient treatment stratification." In: *International journal of cancer* 138.7 (2016), pp. 1765–1776.

Lara Schneider, **Tim Kehl**, Kristina Thedinga, Nadja Liddy Grammes, Christina Backes, Christopher Mohr, Benjamin Schubert, Kerstin Lenhof, Nico Gerstner, Andreas Daniel Hartkopf, et al. "ClinOmicsTrailbc: a visual analytics tool for breast cancer treatment stratification." In: *Bioinformatics* 35.24 (2019), pp. 5171–5181.

Daniel Stöckel, Oliver Müller, **Tim Kehl**, Andreas Gerasch, Christina Backes, Alexander Rurainski, Andreas Keller, Michael Kaufmann, and Hans-Peter Lenhof. "NetworkTrail—a web service for identifying and visualizing deregulated subnetworks." In: *Bioinformatics* 29.13 (2013), pp. 1702–1703.

Daniel Stöckel, **Tim Kehl**, Patrick Trampert, Lara Schneider, Christina Backes, Nicole Ludwig, Andreas Gerasch, Michael Kaufmann, Manfred Gessler, Norbert Graf, et al. "Multi-omics enrichment analysis using the GeneTrail2 web service." In: *Bioinformatics* 32.10 (2016), pp. 1502–1508.

Romina Vardapour, **Tim Kehl**, Susanne Kneitz, Nicole Ludwig, Eckart Meese, Hans-Peter Lenhof, and Manfred Gessler. "The DGCR8 E518K mutation found in Wilms tumors leads to a partial miRNA processing defect that alters gene expression patterns and biological processes." In: *Carcinogenesis* (2021f).

B

ADDITIONAL METHODS

In this chapter additional methods are described that are used in this thesis, i.e., clustering methods (cf. Section B.1) and dimension reduction techniques (cf. Section B.2).

B.1 CLUSTERING

Clustering is a common but very important task in the analysis of many high-throughput data sets and especially large single-cell data sets. Due to the size of the data sets alone, it is often required to partition samples (or features) into groups, which can then be compared.

In this section, we introduce different clustering techniques that are relevant for the work described in this thesis.

B.1.1 *Hierarchical clustering*

Hierarchical clustering is a class of algorithms that build a hierarchy of clusters using greedy approaches. In general there are two iterative strategies: agglomerative (bottom-up) and divisive (top-down) [387]. Agglomerative strategies start with each data point in an individual cluster and in each step try to merge the most similar (or least distant) ones. In contrast, divisive strategies start with all data points in one cluster and in each iteration identify the best split.

In the following, we focus on the agglomerative strategy [387].

B.1.1.1 *Agglomerative hierarchical clustering*

The individual clustering methods for the agglomerative strategy can be distinguished by the way they determine which two clusters should be merged. To this end, they use so called linkage criteria. In general, these measures define the distance of two clusters based on the distances of the contained points. For this purpose, any distance or similarity measure can be applied, e.g., all methods presented in Section 3.5. Examples for the linkage of two clusters using different criteria are depicted in Figure 98 A.
In the following paragraphs, we introduce some of the measures that can be used as linkage criteria. For simplicity, all definitions are based on distance metrics. However, equivalent formulations can be derived for similarity measures.

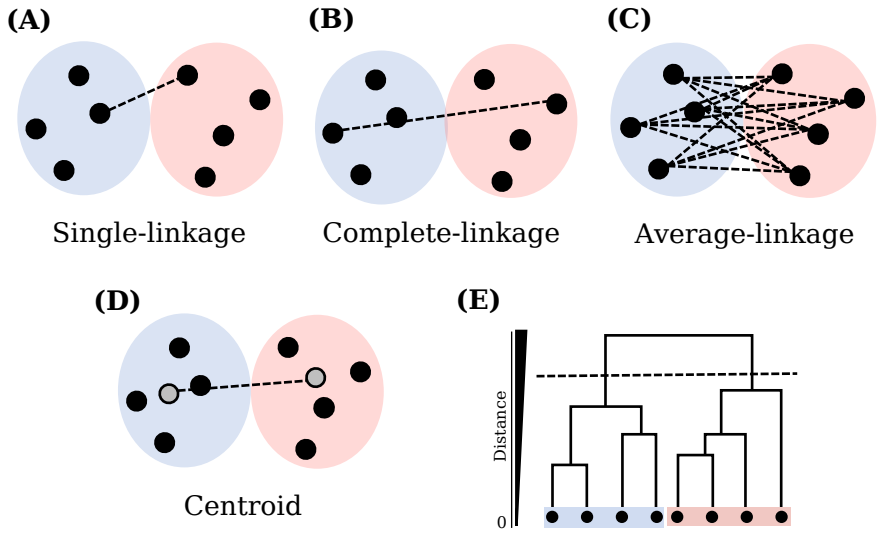


Figure 98: Overview of linkage criteria for hierarchical clustering and dendrogram. Examples for (A) single-linkage, (B) complete-linkage, (C) average-linkage, and (D) centroid-based linkage distance between two clusters. (E) Dendrogram of a single-linkage clustering. The height of the dendrogram indicates the distance between the points. The dotted line marks the threshold used to create the two marked clusters.

B.1.1.2 *Point-based linkage criteria*

Many linkage criteria define the distance between two clusters A and B based on the pairwise distance between all points. Popular functions in this group are:

Single-linkage

The single-linkage strategy defines the distance between two clusters A and B as the minimal pairwise distance between members of cluster A and members of cluster B (cf. Figure 98 A).

$$D(A, B) = \min_{a \in A, b \in B} \{d(a, b)\} \tag{121}$$

Complete-linkage

The complete-linkage measure determines the distance between two clusters as maximum pairwise distance between members of cluster A and members of cluster B (cf. Figure 98 B).

$$D(A, B) = \max_{a \in A, b \in B} \{d(a, b)\} \tag{122}$$

Average-linkage

The average-linkage criterion defines the distance between two clusters A and B as the average distance between members of cluster A and members of cluster B (cf. Figure 98 C).

$$D(A,B) = \frac{1}{|A| \times |B|} \sum_{a \in A, b \in B} d(a,b) \tag{123}$$

B.1.1.3 *Centroid-based linkage criteria*

In contrast to the previously described measures that determine the distance between two clusters based on the contained points directly, other approaches define the distance between two clusters using the centroids of each cluster. The centroid of a cluster is a new point defined by the average position of all member genes (cf. Figure 98 D).

$$D(A,B) = d(\bar{a},\bar{b}) \text{ } \textit{(centroid method)} \tag{124}$$

or

$$D(A,B) = \frac{d(\bar{a},\bar{b})^2}{\frac{1}{|A|} + \frac{1}{|B|}} \text{ } \textit{(Ward's method [564])} \tag{125}$$

Here $\bar{a}$ and $\bar{b}$ are defined as the centroids of cluster A and B respectively.

B.1.1.4 *General agglomerative clustering algorithm*

The agglomerative clustering approach starts with each point in its own cluster. Then, an iterative greedy approach is used that in each step merges the two clusters with the smallest distance according to the selected linkage strategy. This process is repeated until all points are merged. The created cluster hierarchy is often illustrated as a dendrogram depicting the order in which the clusters are merged and which distance they had (cf. Figure 98 E). The dendrogram is then cut at a certain level to retrieve a final cluster composition. The runtime for this approach is in $O(n^3)$, where n is the number of points. The pseudocode for this strategy is depicted in Algorithm 2.

Algorithm 2 Agglomerative hierarchical clustering

Given n initial clusters $C = \{C_1, ..., C_n\}$ *and a*
linkage-criterion $D(C_i, C_j)$.
procedure HCLUST
  while $|C| > 1$ **do**
    //Find cluster pair with minimal distance
    $C_i, C_j = \underset{\forall (C_i, C_j): C_i \neq C_j}{\arg\min} \{D(C_i, C_j)\}$

    //Update clusters
    $C = C \setminus C_i$
    $C = C \setminus C_j$
    $C = C \cup \{C_i, C_j\}$

    //Save the cluster composition and distance to build dendrogram
    save(C, C_i, C_j,D(C_i, C_j))

B.1.2 *Community clustering*

Another class of clustering algorithms are community detection methods, such as Louvain [61] and Leiden [537]. Especially in the context of single-cell analysis they are widely used [320].

In these approaches, the similarity between the different points is modelled as a weighted graph $G = (E, V)$, where each data point represents a node and weighted edges the similarity between them. The clusters are defined as local subgraphs, i.e., communities, in which the member nodes have a higher similarity to other members than to nodes outside the community. For the detection of these communities, a modularity score Q is optimized that compares the sum of edge weights in each community to the remaining edges in the graph. An overview of this approach is depicted in Figure 99.
Given an undirected weighted graph $G = (E, V)$ with n nodes and a similarity matrix $A^{n \times n}$, which for each pair of nodes (v_i, v_j) contains the edge weight A_{ij}. Let additionally k_i be the total edge weights for node v_i and $2m$ the sum of all edge weights in G. Then the modularity is defined as:

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j), \tag{126}$$

where c_i and c_j are the assigned communities of nodes v_i and v_j respectively, and $\delta(c_i, c_j)$ is the Kronecker delta function:

$$\delta(c_i, c_j) = \begin{cases} 1, \text{ if } c_i = c_j \\ 0, \text{ if } c_i \neq c_j \end{cases} \tag{127}$$

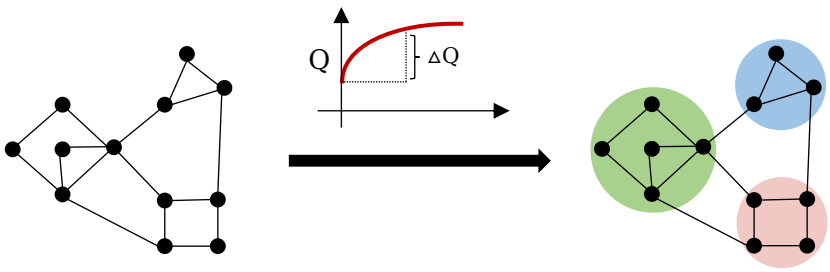


Figure 99: Overview of the community clustering strategy. In this approach, all points are interpreted as nodes in a graph and edges between them are weighted with the similarity between two points. For simplicity, only edges above a certain similarity threshold are depicted. The different clustering algorithms then try to identify clusters as communities in the graph by optimizing the modularity function Q.

Both the Louvain and the Leiden method use greedy approaches that try to maximize Q. In the following paragraphs, both methods are briefly summarized.

B.1.2.1 *Louvain*

The Louvain algorithm [61] usually starts with each vertex in its own cluster. Then, the following two-step approach is applied iteratively to optimize Q.

Step 1: local community optimization

In the first step, all nodes are processed in an arbitrary order to optimize the local modularity score. For each node v_i and each community c_n that contains a neighbour nodes of vi, we calculate how the modularity score would change if we remove v from it current community and add it to c_n. The node v is then assigned to the neighboring community that leads to the largest increase in modularity. If none of these assignments leads to an increase in Q the node stays in its original cluster.
After all nodes have been processed sequentially, this procedure is repeated in the same order until a local maximum of Q is reached.

According to the authors, the order in which the nodes are processed influences the outcome only slightly and mainly affects the runtime of the algorithm [61].

Step 2 - network aggregation

In the second step, we use the community assignment obtained in Step 1 to build a new graph. To this end, nodes in the same commu-

nity are merged into one node in the new graph. All edges between nodes in the same community are converted into self edges, and all edges between nodes of two different communities remain weighted edges between the corresponding nodes in the new graph.
After the newly build graph is completed, Step 1 is repeated for the aggregated network.

B.1.2.2 *Leiden*

The Leiden algorithm by Traag et al. is an adaptation of the Louvain algorithm, which ensures that all found communities are well connected, e.g., that a community does not contain any disconnected components [537]. This is achieved by introducing a new step in between the local community optimization and the network aggregation, which refines the community assignment [537].
On top of this, the Leiden algorithm also uses an adapted approach for the local community optimization that only reiterates over nodes that have been assigned to new communities. This leads to an improved runtime [537].

B.2 DIMENSION REDUCTION

The visualization of molecular high-throughput data sets is a crucial, but often challenging task. The high dimensionality of the data makes it hard to get an overview or to highlight important aspects.
For this purpose, researchers often rely on dimension reduction techniques that project the data into two or three dimensions, which are then easier to visualize (cf. Figure 100).
In the following paragraphs, we present an overview of the most popular techniques.

B.2.1 *Principal component analysis (PCA)*

A principle components analysis (PCA) is a transformation that projects data points into a subspace with lower dimension, while retaining as much information as possible. In this context, the principle components are a series of uncorrelated, orthogonal projections ordered by variance [163]. The principle components constitute an orthonormal basis that can be utilized to conduct a change of basis.

Given a matrix $X^{n\times p}$ with n rows and p columns, the PCA can be formulated as a minimization problem, in which we find the best fitting orthogonal matrix $V_q^{n\times q}$ that satisfies [163]:

$$\min_{V_q} \sum_{i=1}^{n} \|(x_i - \bar{x})V_q V_q^T (x_i - \bar{x})\|^2 \tag{128}$$

Using V_q, we can project any (centered) point $\tilde{x_i} = x_i - \bar{x}$ into a q-dimensional subspace ($V_q V_q^T \tilde{x_i}$) defined by the columns of V_q.
Solutions for this problem can be obtained in various ways, e.g., using singular value decomposition [163] or autoencoder neural networks [128].

B.2.2 *t-distributed Stochastic Neighbor Embedding (t-SNE)*

$t-SNE$ is a non-linear dimensional reduction technique that was designed to visualize high-dimensional data sets in two or three dimensions [324]. In this approach, the original high-dimensional representation of the data is converted into a low-dimensional one that retains as much of the distance information as possible.
To this end, both representations are modelled as neighborhood graphs, i.e., undirected and weighted graphs, in which each vertex represents a data point and each edge the similarity between two points. Here, each data point belongs to one specific vertex in both graphs. To obtain the best dimensional reduction, the positions of the vertices in the low dimensional representation are then optimized such that they reflect the high-dimensional structure of the data.

Given n data points $x_1, ..., x_n$ in high-dimensional space and corresponding data points in low-dimensional space $y_1, ..., y_n$. The similarity of a point x_i and a point x_j in high-dimensional space is defined using a gaussian kernel that is centered around x_i:

$$v_{i|j} = e^{\frac{-||x_i - x_j||^2}{2\sigma_i^2}}, \tag{129}$$

where σ_i is the standard deviation of the kernel. How the value for σ_i can be determined, is discussed later in this section.
For each $v_{i|j}$, we can now define a probability $p_{i|j}$:

$$p_{i|j} = \frac{v_{i|j}}{\sum_{k \neq i} v_{i|k}} \tag{130}$$

$p_{i|j}$ determines how likely it is that x_j is selected as a neighbor of x_i compared to all other points. Finally this probability is symmetrized and normalized over all data points:

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}, \tag{131}$$

As indicated previously, each $p_{i|j}$ depends on the standard deviation σ_i, and each value of σ_i defines a different probability distribution P_i that is centered around x_i. Accordingly, the value σ_i influences the number of effective neighbors ($v_{ij} \gg 0$) that are considered. Hence, their values need to be carefully selected.
In the t-SNE algorithm, the value for each σ_i is selected via a binary

search that ensures the following constraint about a user defined parameter "Perplexity" is satisfied:

$$\text{Perplexity}(P_i) = 2^{-\sum_j p_{j|i} \log_2(p_{j|i})} \tag{132}$$

Intuitively, the perplexity is a measure that loosely defines the number of effective neighbors that should be considered for each point.

Next, we can define the low-dimensional embedding. Here, the similarity of two points y_i and y_j is modelled accordingly, however, using a Student t-distribution:

$$q_{ij} = \frac{w_{ik}}{\sum_{k \neq i} w_{ik}} \tag{133}$$

$$w_{ij} = (1 + \|y_i - y_j\|_2^2)^{-1} \tag{134}$$

Finally, the locations for the points in a two- or three-dimensional embedding are optimized by a gradient descent algorithm that uses the Kullback–Leibler divergence between the two probabilities p_{ij} and q_{ij} as a loss function:

$$KL(P\|Q) = \sum_{i \neq j} p_{ij} \log_2(\frac{p_{ij}}{q_{ij}}) \tag{135}$$

B.2.3 *Uniform Manifold Approximation and Projection (UMAP)*

Uniform Manifold Approximation and Projection (UMAP) [345] is an alternative method for dimension reduction that uses a similar strategy as $t-SNE$. First, it models the data as two neighborhood graphs. One in high-dimensional and one in low-dimensional space. Then, the coordinates of all points in the low-dimensional space are optimized such that is reflected the high-dimensional structure of the data. It can be defined analogously to $t-SNE$ [345]:
Given n data points $x_1, ..., x_n$ in high-dimensional space and corresponding data points $y_1, ..., y_n$ in low-dimensional space. The similarity of a point x_i and a point x_j in high-dimensional space is defined as:

$$v_{i|j} = e^{-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}}, \tag{136}$$

where d is an arbitrary distance function, ρ_i the distance to the closest neighbor of x_i and σ_i a normalization factor. In the UMAP algorithm $v_{i|j}$ is only calculated for the k closest neighbors of i and all others are set to 0 [345]. The value for k is a user-defined parameter.
Next, the similarities are symmetrized as follows[345]:

$$v_{ij} = v_{i|j} + v_{j|i} - v_{i|j} \cdot v_{j|i} \tag{137}$$

Like in the t − SNE approach, the value of each $v_{i|j}$ depends on the normalization factor σ_i. Similarly, they are selected using a binary search that ensures the following constraint about the number of neighbors k is satisfied:

$$\log_2(k) = -\sum_{i=1}^{n} v_{ij} \tag{138}$$

The similarities for two points y_i and y_j in low-dimensional space are defined as follows:

$$w_{ij} = (1 + a \cdot \|y_i - y_j\|_2^{2b})^{-1}, \tag{139}$$

where a and b are positive values defined by the user.
The positions of all points in the low-dimensional embedding are then optimized by a stochastic gradient descent algorithm that uses the cross-entropy loss function:

$$H(V, W) = \sum_{i \neq j} v_{ij} \cdot \log_2(\frac{v_{ij}}{w_{ij}}) - (1 - v_{ij}) \cdot \log_2(\frac{1 - v_{ij}}{1 - w_{ij}}) \tag{140}$$

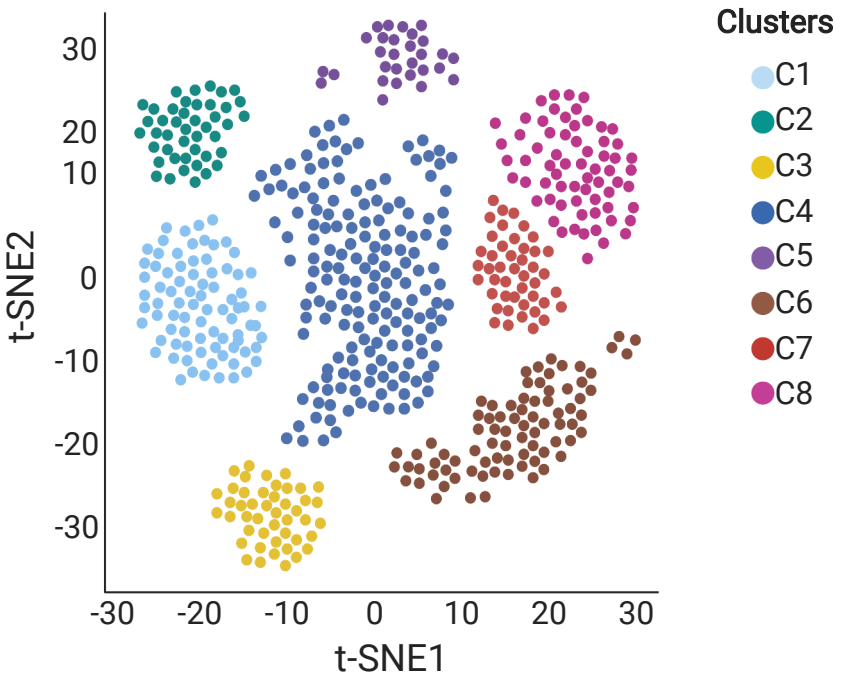


Figure 100: Example of a scatterplot depicting 2D coordinates of a tSNE dimension reduction. The data points are colored according to a cluster assignment. This figure was created using BioRender.com.

C

TOOLS FOR CLINICAL DECISION SUPPORT

Cancer is a class of diseases that are characterized by cells that have the ability to proliferate uncontrollably and the potential to invade surrounding tissues. In 2019, around 231,000 people in Germany died of cancer, which makes it the second leading cause of death overall and the most frequent cause of death for people with an age between 45 to 65 [393]. While extensive medical research over the last decades has significantly improved the treatment of cancer, the selection of suitable therapy options for a patient is still a major challenge. One reason for this is the high heterogeneity of cells within a tumor, which is enabled by hallmark characteristics, like genetic instability and high proliferation rates that drive the malignancy of the disease [337] (cf. Section 2.3). The genetic diversity of tumors cells can also often lead to phenotypic differences that can impact therapeutic outcomes [337]. For example, it has been shown that specific molecular events can heavily affect the efficiency of certain treatments [153], e.g., activating mutations of the KRAS proto-oncogene in colorectal cancer can potentially cause an resistance to the drug cetuximab [307].
Since high-throughput techniques have the potential to identify such molecular markers, one major goal in current clinical research is to evaluate if the incorporation of these technologies into clinical decision-making can improve assessment of therapy options [308, 452].
As a result, cancer centers started to implement molecular tumor boards (MTBs), where an interdisciplinary team of medical professionals evaluates the best possible therapy options for cancer patients. In addition to clinical information and the patient's treatment history, these MTBs already consider genetic aberrations for the treatment stratification process [549].
Accordingly, over the last few years, different tools have been proposed that combine clinical information of patients with molecular measurements to support clinicians and molecular tumor boards in their quest to find the best possible care for a patient. These include commercially available testing kits, tools, and services that evaluate treatment options based on clinical characteristics and molecular measurements, e.g., offered by Agendia BV (MammaPrint) [491], CeGaT [180], Foundation Medicine (FoundationOne CDx) [349], or Genomic Health (Oncotype DX) [403]. Additionally, various computational tools have been developed that analyze clinical information of molecular measurements to predict the survival time of patients or to evaluate specific therapy options, e.g., Adjuvant! Online [58], CancerMath [359], and PREDICT [126]. A comparison of tools

and services for this purpose is shown in Figure 101.
Here, we present DrugTargetInspector and ClinOmicsTrail[bc], two web services for clinical cancer decision support. Both tools are designed to help clinicians with the assessment of potential treatment strategies based on the molecular characteristics of a tumor. To this end, they provide rich functionality for the integrative analysis of different high-throughput measurements of a cancer sample. These measurements can help to identify important molecular characteristics of an investigated tumor that might help to assess if certain therapeutic approaches might be applicable.

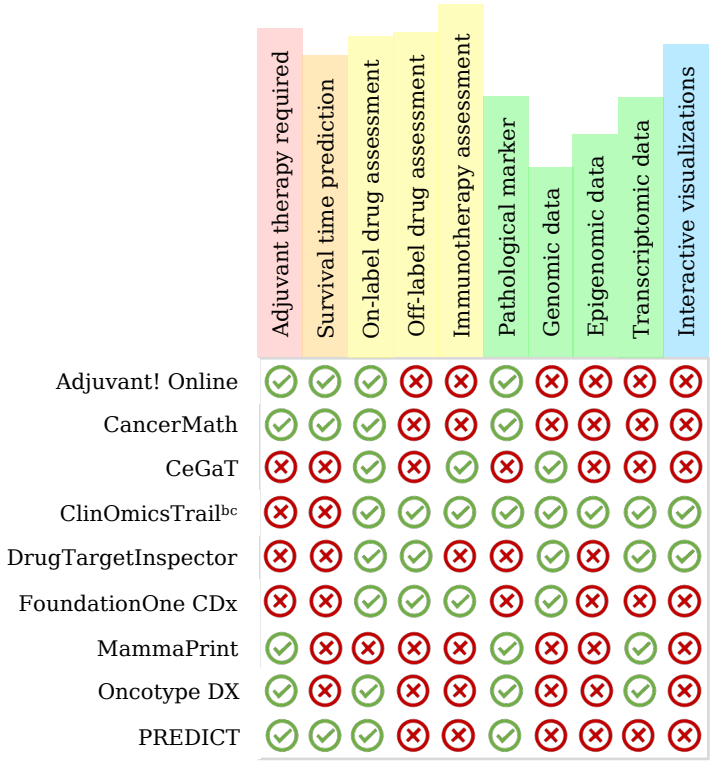


Figure 101: Overview of tools and services for clinical decision support. Green checkmarks indicate that this feature is supported by the tool, and red crosses that the feature is not available. This table is adapted from [475].

C.1 DRUGTARGETINSPECTOR (DTI)

Author contributions

This section is based on our publication "DrugTargetInspector: An assistance tool for patient treatment stratification" [474]. DrugTargetInspector was mainly developed by Lara Schneider and Hans-Peter Lenhof. I was involved in the website design, integration of downstream functionality, and maintenance of the tool. The complete list of contributors can be found in the author section of the respective publication [474].

DrugTargetInspector (DTI) is an assistance tool for cancer treatment stratification. Our web service analyzes multi-omics measurements of a tumor to assess if molecular drug targets are altered, how these alterations might affect down-stream processes, and if this could impact the efficacy of the respective drug. To this end, our web service offers functionality for the identification of deregulated or mutated drug targets, associated pharmacogenomic effects, and their influence on biological processes. A general overview of the DTI workflow is presented in Figure 102.

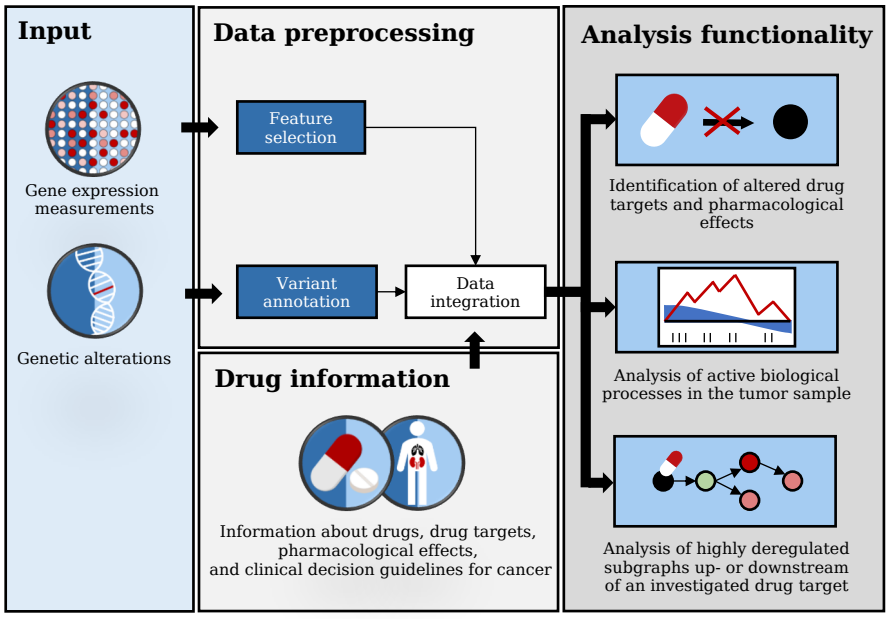


Figure 102: Overview of the DrugTargetInspector workflow. The different input data types are processed individually and then integrated with drug information from external resources. From here, different computational analyses can be performed.

C.1.1 *Identification of altered drug targets and their impact on biological processes*

The inputs for DrugTargetInspector are mutations and gene expression measurements for a tumor sample and a corresponding control or panel of controls. The provided measurements are then used to investigate altered drug targets and associated biological processes. To this end, the data types are first processed individually and then combined in an integrative view that provides an in-depth characterization of a tumor's genetic and molecular profile (cf. Figure 103A). In the following paragraphs, we first describe the processing steps for the input data and then discuss how DTI uses the provided information to evaluate treatment options.

C.1.1.1 *Data processing and integration*

The gene expression data can be uploaded as a matrix containing normalized expression measurements for both tumor sample and corresponding control(s) (cf. Section F.1.1.3). After the file upload, the two groups are compared to identify differentially expressed genes. To this end, users select any of the methods available in GeneTrail (cf. Chapter 5). Alternatively, users can directly upload a score list to our web server.
The mutation data can be provided as a text file in VCF format, see Section F.2.3. The uploaded genetic variants are then processed using the Ensembl Variant Effect Predictor (VEP) [346]. Amongst others, this step identifies genes and proteins that are affected by a particular alteration. Additionally, each variant is annotated with its consequences on the protein sequence, e.g., frameshift, missense, or start lost. Based on this information, the variants are filtered and only protein-altering variants are retained for further processing.
After gene expression measurements and genetic variants are processed, they are used to identify all drug targets that contain mutations or that are differentially expressed between the tumor and associated controls. For each drug with a target in the resulting list, we also compile additional information from our data collection (cf. Section 3.2.6), such as links to external databases or known pharmacogenomic events, i.e., alterations of molecular features that influence the effectiveness of the considered drug.

C.1.1.2 *Identification of altered drug targets and pharmacogenomic effects*

After data processing, we obtain a compact interactive table that contains all genes that are targeted by at least one drug (cf. Figure 103A). Here, users can select if the drug information should be restricted to recommended drugs for the investigated cancer type (on-label drugs) or contain all available information (on-label and off-label drugs).

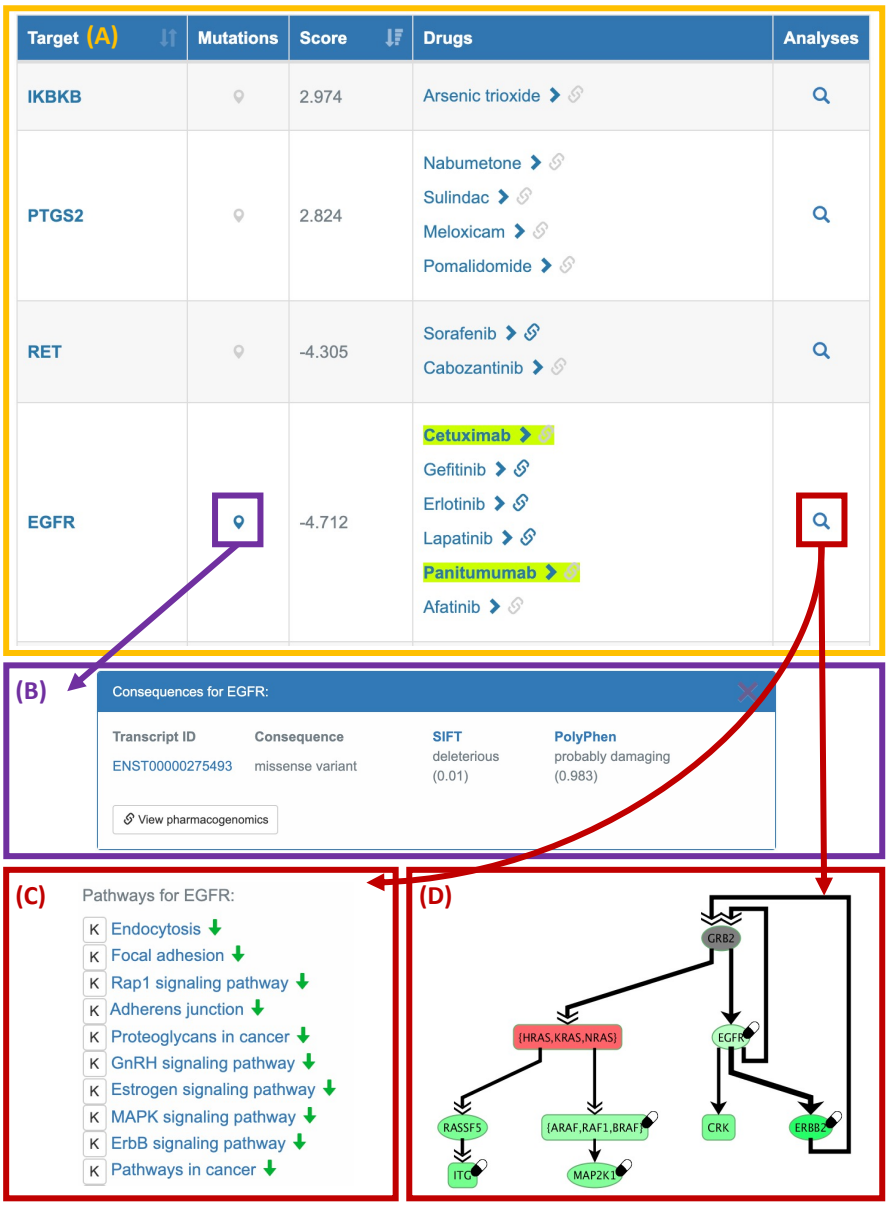


Figure 103: Overview of the different results provided by DrugTargetInspector for the colon adenocarcinoma sample TCGA-AA-3542.
(A) Table with differentially expressed or mutated drug targets. Recommended drugs for colon cancer are highlighted in green. (B) Predicted consequences of the identified EGFR mutation on the protein sequence. (C) Summary of enrichment biological categories that contain EGFR. (D) BiNA [175] visualization of the most deregulated subgraph downstream of EGFR.

Each line in the resulting table depicts information about a specific drug target, all protein-altering mutations within the respective gene, and all drugs that target this molecule. From here, users have different options. On the one hand, they can obtain additional information about the identified mutations or the considered drugs. For mutations, DTI offers a compact table that summarizes the consequence the mutation has on the respective gene and different measures that assess the severity of the alteration, i.e., SIFT [285] and PolyPhen [6] (cf. Figure 103B). For drugs, DTI provides pharmacogenomic information from the Genomics of Drug Sensitivity in Cancer (GDSC) project [594]. This database contains a list of alterations in molecular features that affect the efficacy of cancer drugs, such as mutations in specific genes cause a resistance to a certain drug. Hence, the provided information can be used to assess if the mode of action of individual drugs could be impeded and should not be part of the treatment regimen.

C.1.1.3 *Analysis of altered target pathways*

The research in the last decades has shown that an increased activity of several key signaling pathways is directly connected with the malignancy of various cancer types [339, 368, 430]. For example, elevated levels of HIF-1α pathway seem to be associated with tumor angiogenesis, tumor metastasis, and resistance to cancer therapy [339]. Consequently, central genes those pathways are extensively studied as novel targets for anti-cancer agents.
Hence, our web service also provides functionality for the analysis of altered signaling pathways that are associated with a specific drug target. To this end, DrugTargetInspector offers two different analyses that presented in the following paragraphs.

Analysis of enriched biological processes

Starting with a drug target of interest, our web service can be applied to examine which signaling pathways might be affected by the corresponding gene. To this end, we conducted Gene Set Enrichment Analyses for all KEGG pathways [394] that contain the analyzed target. In order to inspect these results, DTI offers two ways. An overview of the most significant categories that involve the examined target genes is directly shown on the main DTI result page (cf. Figure 103C). Additionally, a detailed view of the results is provided. To this end, users are redirected to the GeneTrail result visualization (cf. Figure 53).
The described enrichment analysis functionality is especially useful to determine if the target pathway of a particular drug is active or inactive in the investigated tumor sample. This information provides evidence if the drug's mode of action might be impeded and could affect a putative therapy.

Detection of deregulated signaling cascades

In addition to the pathway activity patterns, DTI can also be applied to study up- and downstream effects of a specific drug target in the KEGG network [394]. To this end, we use a rooted version of the subgraph ILP described in Section 3.7.1. For this purpose, we extended the ILP with an additional constraint that fixes the investigated drug target as the root in the selected subnetwork. Starting from this root, the ILP identifies the most deregulated subgraph in the analyzed network that is located directly upstream or downstream of the specified root. For both tasks, we use the same approach. However, for the former, we reverse the edge directions in the optimization problem.

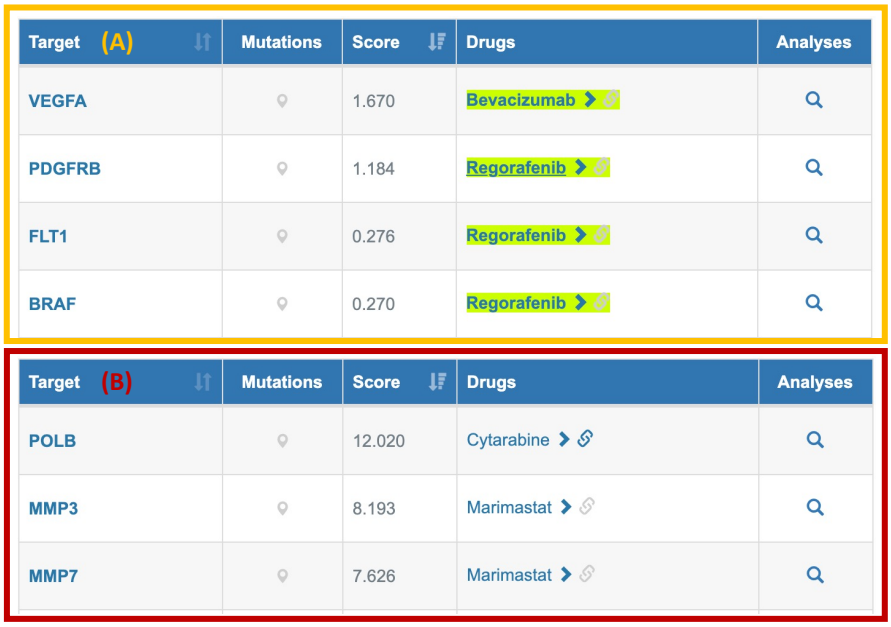
C.1.2 *Example application: Colon adenocarcinoma (TCGA-AA-3542)*

In order to demonstrate the capabilities of our tool, we analyzed genetic variations and gene expression measurements of a colon adenocarcinoma sample obtained from The Cancer Genome Atlas (TCGA-AA-3542). In particular, we used DrugTargetInspector to evaluate and compare different targeted therapy options for colon cancer. First, we discuss drugs that are recommended for colon cancer via clinical decision guidelines provided by the European Society for Medical Oncology (ESMO). These are Bevacizumab, Cetuximab, Panitumumab, and Regorafenib.

For our analysis, we first uploaded gene expression measurements for the tumor sample and a panel of nine control samples. To compare both groups, we calculated a Z-score for each gene. In the second step, we uploaded genetic variants in VCF format. These alterations were then processed using VEP and subsequently filtered, such that only protein-altering variants remain. The complete list of parameters can be found in Appendix E.10. An excerpt of the result table is shown in Figure 103A.

As depicted, we see that the epidermal growth factor receptor (EGFR), the target molecule for both Cetuximab and Panitumumab, is highly downregulated in the tumor and additionally contains a missense mutation (Figure 103B). Both Cetuximab and Panitumumab are antibodies that inhibit EGFR. Since we assume that inhibitors are more effective of their target is highly expressed, a downregulation of EGFR in the tumor suggests that a treatment with either one of the drugs could be compromised. Next, we conducted a subgraph analysis based on the KEGG network with EGFR as the root. The resulting subnetwork is shown in Figure 103D. Here, we observe that many genes that are located downstream of EGFR in the KEGG network are also downregulated. This provides further evidence that a therapy with either Cetuximab and Panitumumab might be ineffective for the investigated tumor.

For Bevacizumab, we observe that its molecular target VEGFA is upregulated in the tumor and not mutated. This suggests that this drug might work as expected and could potentially be considered to be part of a treatment regimen (cf. Figure 104A). We obtain similar results for some target molecules of Regorafenib. However this kinase inhibitor has several targets genes, some of which are also downregulated, which could potentially impede its efficacy. Hence, based on our analysis Bevacizumab seems to be the best candidate amongst the four recommended drugs for colon cancer.



Target (A)	Mutations	Score	Drugs	Analyses
VEGFA		1.670	Bevacizumab	
PDGFRB		1.184	Regorafenib	
FLT1		0.276	Regorafenib	
BRAF		0.270	Regorafenib	

Target (B)	Mutations	Score	Drugs	Analyses
POLB		12.020	Cytarabine	
MMP3		8.193	Marimastat	
MMP7		7.626	Marimastat	

Figure 104: Overview of DrugTargetInspector results for different targeted drugs. (A) Target molecules for Bevacizumab and Regorafenib. For Regorafenib only upregulated target genes are shown. (B) The most upregulated target molecules in our DTI analysis. The corresponding drugs Cytarabine and Marimastat could be interesting options for an off-label use case.

Besides the recommended drugs, we also used DTI to assess putative off-label drugs that could be considered. Figure 104A depicts the most upregulated target genes in our analysis and the corresponding drugs Cytarabine and Marimastat. Cytarabine is an antimetabolic agent with on-label use for different types of leukemia [137]. Additionally, it can be used as off-label drug for non-Hodgkin's lymphoma and primary central nervous system (CNS) lymphoma [137]. Hence it is most likely not a suitable option for the analyzed colon adenocarcinoma. The second candidate, Marimastat is an antimetastatic agent that targets matrix metalloproteinases, like MMP3 and MMP7. So far various clinical trials have been conducted that analyzed the efficacy of this drug in cancer patients [483, 617]. However,

a treatment with Marimastat did not improve the survival time of the respective patients and even had a negative influence on their quality of life. Hence, this drug would also not be a suitable option for the investigated tumor.

C.2 CLINOMICSTRAIL^{bc}

Author contributions

This section is based on our publication "ClinOmicsTrail^{bc}: a visual analytics tool for breast cancer treatment stratification" [475]. ClinOmicsTrail^{bc} was mainly designed by Lara Schneider and Hans-Peter Lenhof. I was involved in the implementation and data integration of the web service. The complete list of contributors can be found in the author section of the respective publication [475].

ClinOmicsTrail^{bc} is an assistance tool for clinicians or molecular tumor boards that can be used to assess putative treatment options for breast cancer patients based on clinical markers and molecular characteristics of a tumor. To this end, our web service analyzes molecular high-throughput profiles of this tumor and corresponding controls to identify (epi-)genetic alterations in tumor biomarkers, such as driver genes, relevant cancer pathways, or molecular drug targets. These alterations can then be used to assess the suitability of different treatment options, i.e., on- and off-label drugs, or immunotherapeutic approaches.

C.2.1 *Input data and initial processing steps*

The inputs for ClinOmicsTrail^{bc} are molecular profiles and clinical information of the considered tumor and at least one associated control sample. The controls can either be healthy samples or other tumor samples that can help to identify important molecular traits of the sample of interest. In general, ClinOmicsTrail^{bc} can be applied to analyze gene expression profiles, genetic variants, copy number alterations, and cytosine methylation patterns. While all data types provide valuable insights into cellular mechanisms and can help in the treatment selection process, only gene expression values are required to start an analysis using our toolbox.
In the following, we briefly discuss the different input types and the initial processing steps conducted by our tool.

C.2.1.1 *Clinical markers*

Clinical markers of the tumor under investigations are some of the most important features used to select a suitable therapy for a breast

cancer patient. These markers include patient-specific information such as the menopausal status, and tumor-specific markers like the grade, size, sample origin, lymph node status, metastasis status, presence of hormone receptors on the surface, the histopathological subtype, amplification of the HER2 receptor, and information about the growth rate (Ki-67 staining, S-phase fraction).
If available, the clinical markers can directly be entered into the web interface.

C.2.1.2 *Gene expression values*

Gene expression values can either be uploaded as a score list or a matrix containing normalized expression values for both the tumor and matched controls. For matrices, our web service applies a group comparison step to calculate gene expression differences. To this end, either fold-changes or Z-scores are applied for all genes.

C.2.1.3 *Copy number variations (CNVs)*

In breast cancer, copy number variations (CNVs) seem to be responsible for up to 85% of deregulated genes [506]. Hence, it is helpful to directly incorporate CNVs into a ClinOmicsTrail^{bc} analysis. To this end, users can upload CNVs for genomic regions in SEG format that contain log-ratios between copy numbers in the tumor and normal tissues (cf. Section F.2.4). In our web service, the uploaded regions that overlap with gene annotations are aggregated for each gene using BEDTools [426].

C.2.1.4 *Genetic alterations*

During their development, cancer cells often accumulate a large number of genetic alterations that enable or facilitate the acquisition of cancer hallmarks (cf. Section 2.3). Additionally, certain mutations are also known to affect cancer treatment. Here, in particular, alterations in target molecules of a particular drug can heavily influence their effectiveness.
To incorporate genetic variations in a ClinOmicsTrail^{bc} analysis, users can upload them as text file in VCF format (cf. Section F.2.3). The uploaded genetic variants are then processed using the Ensembl Variant Effect Predictor (VEP) [346] and filtered to identify all protein altering mutations. The resulting variants are further annotated using external databases from our data collection (cf. Section 3.2.6), such as pharmacogenomic information from the GDSC database [238], or cancer driver information from the IntOGen database [185]

C.2.1.5 *DNA methylation patterns*

Further factors contributing to tumor initiation and progression are epigenetic changes, such as DNA methylation patterns.
Similar to gene expression values, these can be provided as text files containing pre-computed scores or normalized measurements for the tumor and matched controls. Here, we assume that all values are already aggregated on a gene level. If multiple samples are provided, our web service automatically conducts a group comparison using either fold-changes or Z-scores.

C.2.2 *Identification of tumor characteristics*

Based on the clinical information and molecular profiles provided, our web service conducts further analyses that highlight molecular characteristics of the investigated tumors and help to assess treatment options.
For all analyses in this context, we focus on essential tumor driver genes relevant for the therapy stratification of breast cancer patients. This enables clinicians to get a comprehensive, yet compact overview of a tumor's molecular makeup.

C.2.2.1 *Overview of tumor driving alterations*

In a first step, ClinOmicsTrail^{bc} combines the information provided by the user to create several intuitive representations that help to analyze the molecular characteristics of an investigated tumor.
To this end, our web server generates several tables that indicate if cancer driver genes, molecular drug targets, or other relevant genes are altered in any of the provided data types. Moreover, an interactive sunburst chart is created that combines all information for relevant driver genes in a circular fashion (cf. Figure 105). In this plot, breast cancer marker genes are arranged in concentric circles that represent different types of information available for the respective genes.

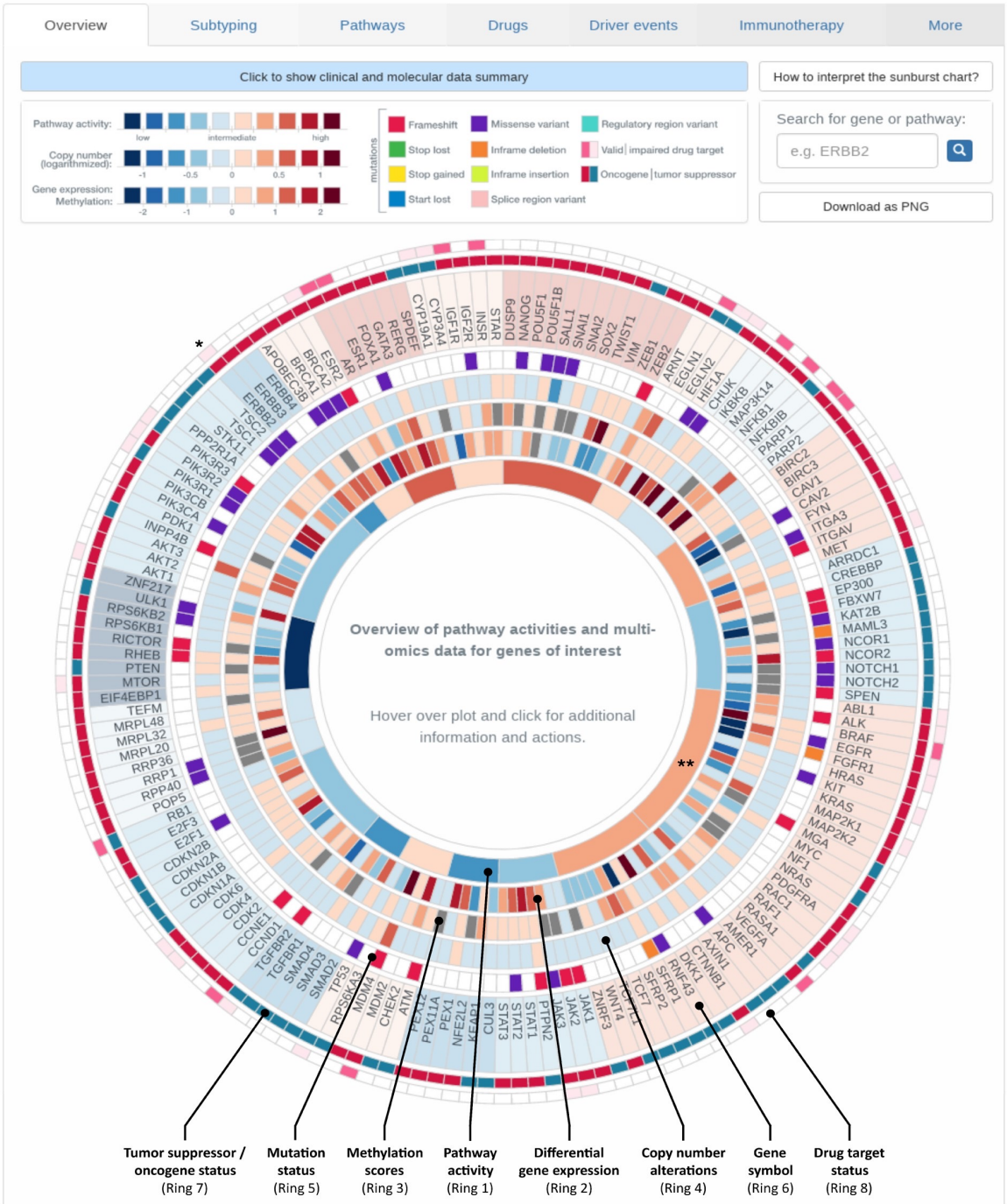


Figure 105: Overview of a tumor sample's molecular characteristics (TCGA-BH-A0DT). In this plot, relevant cancer driver genes are depicted as a sunburst chart. Each slice represents a particular gene and each ring a specific data type. The genes are grouped based on their associated pathway. This figure was retrieved from the original ClinOmicsTrail^{bc} publication [475]

.

C.2.2.2 *Analysis of pathway activities*

Altered signaling pathways are an important factor in the development and progression of cancer and often drive the hallmark characteristics of tumors [462]. Due to their crucial role in oncogenesis, they are becoming increasingly important targets in the therapy of cancer [462, 596]. For example, tumors with elevated activity of the PIK3/AKT/mTOR signaling seem to be more receptive for mTOR inhibitors like Everolimus [417].
For this reason, we try to provide a comprehensive overview of altered signaling pathways in our web service. To this end, we first merged annotations for 20 breast cancer-relevant pathways from the different databases in our data collections, i.e., GO [100], KEGG [394], Reactome [142], and WikiPathways [256].
For the resulting gene sets, we created a custom measure that evaluates the activity of those pathways in a given tumor. Our approach is based on the assumption that a targeted treatment is more effective if the corresponding target pathway is highly active [18]. In order to incorporate this information into our approach, we combine the uploaded gene scores with drug sensitivity information from the GDSC database to estimate pathway activities. A detailed description of the methods can be found in previous publications [475, 476]. In this section, we focus on the different applications of this method in ClinOmicsTrail^{bc}.
In our web service, the calculated pathway activities for a given tumor are used in different ways: they are depicted as the most inner ring in our sunburst chart (cf. Figure 105), they are used to assess targeted therapy options (cf. Section C.2.3), and they can be examined on their own.
For the latter, we created an interactive radar chart, where the pathway activities of the uploaded tumor can be compared against breast cancer cell lines or tumors of the TCGA cohort (cf. Figure. 106). This makes it possible to identify cell lines that show activity patterns similar to the investigated sample. Since we assume that the activity of specific pathways is directly associated with the tumor's response to certain treatments, this information can be used to check which drugs were effective or ineffective in similar reference samples. For this purpose, we list all available clinical marker and pharmacological information of the incorporated reference samples.

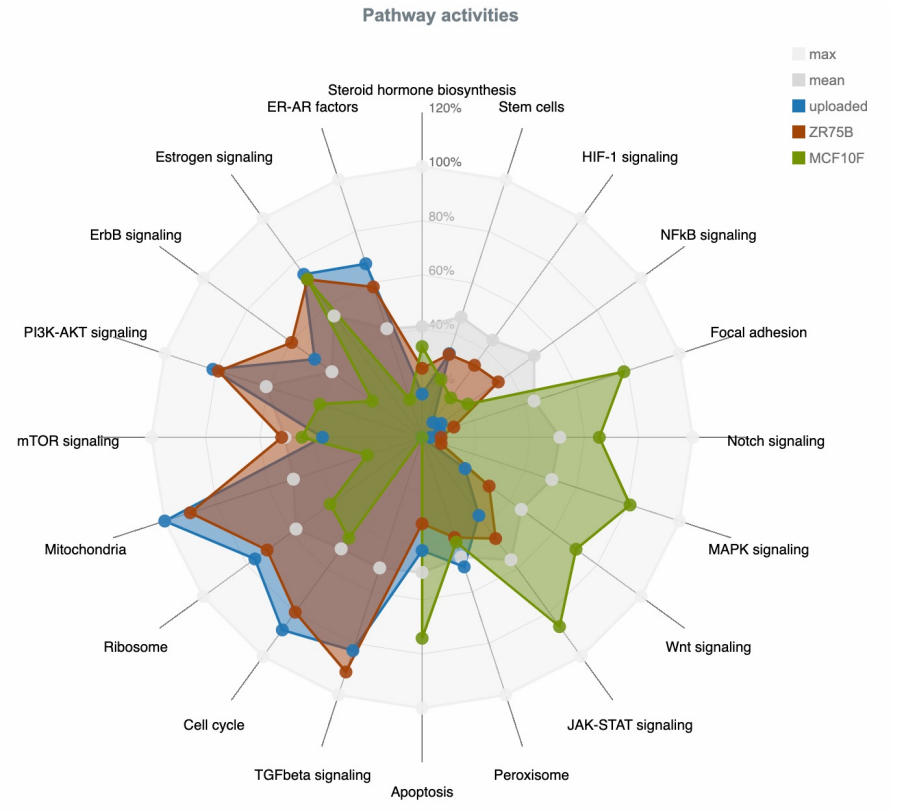


Figure 106: Comparison of pathway activity patterns of three breast cancer cell lines: CAMA1 (uploaded sample, blue), ZRF75B (brown), and MCD10F (green). This figure was retrieved from the original ClinOmicsTrail^{bc} publication [475]

C.2.3 *Assessment of therapy options*

The clinical and molecular characteristics of an investigated tumor cannot only help clinicians and molecular tumor boards to assess the molecular makeup of a tumor, but the provided information can also be utilized to evaluate different treatment options. To this end, ClinOmicsTrail^{bc} integrates and processes the previously described data to extract molecular markers that can influence the efficacy of different therapeutic approaches for breast cancer. Our web service then compiles this information such that clinicians can easily assess if a particular treatment should be considered for a given tumor or if there are any impediments.

C.2.3.1 *Targeted therapy*

In a first analysis, our web service evaluates if the different molecular characteristics of a tumor could potentially influence the efficacy of 17 FDA-approved breast cancer drugs. To this end, ClinOmicsTrail^{bc} analyzes a range clinical and molecular markers for each drug that

indicate if a drug is a suitable therapy option or if its mode of action could be impeded (cf. Figure 107). In this context, our web service first assesses the state of predictive biomarkers, such as the presence of absence of hormone receptors on the surface of the tumor cells that are prerequisite for certain drugs [540]. For example, aromatase inhibitors are usually only administered if the tumor cells are estrogen and progesterone positive [540]. A further factor that can heavily influence the efficacy of a drug are mutations in drug-processing enzymes that can compromise their ability to metabolize the considered drug and, hence, might render a drug inactive. Consequently, ClinOmicsTrail^{bc} also evaluates if drug-processing enzymes contain any genetic alterations that could alter their function. Moreover, our web service assesses if molecular drug targets and associated signaling pathways are active and not affected by any (epi)genetic alteration, which could impede the efficacy of a drug. Finally, ClinOmicsTrail^{bc} also lists alterations in known resistance-promoting factors that could render a drug ineffective.
In summary, this functionality can help clinicians assess the suitability of a specific drug or drug combination and identify potential problems that could interfere with a drug's mode of action.
In addition to the assessment of on-label drugs approved for breast cancer (cf. Figure 107), our web service can also be used to evaluate a collection of "driver targeting drugs", i.e., off-label drugs approved for cancer in general that can be considered if a specific genetic alteration is present.

C.2.3.2 *Immunotherapy*

Besides on- and off-label drugs, ClinOmicsTrail^{bc} can also be applied to assess different cancer immunotherapy options. Here, the goal is to increase the immune system's ability to recognize and destroy cancerous cells [104]. An important feature that is linked to the efficacy of many cancer immunotherapies, e.g., checkpoint inhibitors or antigen vaccination, is the mutational load of a tumor, which is also called the tumor's mutational burden (TMB) [186, 290]. We determine the TMB as the number of somatic mutations that were identified per $1,000,000$ bases in the exome of a given tumor. As a reference, this value is then compared to TMB values for the entire TCGA cohort.
Additionally, our web service can also be used to evaluate the status of other relevant markers for cancer immunotherapies, such as DNA repair genes, members of the human leukocyte antigen (HLA) family, or known biomarkers for checkpoint blockade immunotherapy.
In addition to usage of checkpoint inhibitors, another popular approach for cancer immunotherapy are tumor-specific cancer vaccines [400, 459]. The basis of such vaccines are tumor-specific neoepitopes, i.e., short DNA sequences that distinguish the tumor from respective controls. The neoepitopes can then be used to train T cells that should

be able to recognize these epitopes and the corresponding cancer cells to initiate an immune reactions that destroys them.
For the identification of such neoepitopes, we connected our web service with the ImmunoNodes toolbox [479], which provides several immunoinformatics tools that can be applied for HLA genotyping or the design tumor-specific epitope-based cancer vaccines.

C.2.4 *Example application: Hormone receptor-positive,* HER2*-negative breast tumor (*TCGA-BH-A0DT*)*

In order to demonstrate how ClinOmicsTrail^{bc} can assist in clinical decision making, we are analyzing molecular measurements of a tumor sample from TCGA. The tumor sample was obtained from a 41-year-old woman with a stage II, estrogen and progesterone receptor-positive, HER2-negative breast cancer (TCGA-BH-A0DT).
All processing steps, and the complete set of parameters are described in Appendix E.11. An overview of the results is depicted in Figures 105 and 107.
Based on the clinical markers of the tumor (hormone receptor-positive, HER2-negative), a typical treatment regimen might include the estrogen receptor modifier tamoxifen, which is one of the most commonly prescribed breast cancer drugs [478]. However, around $30-50\%$ of patients with adjuvant tamoxifen treatment exhibit a relapse [478]. One reason for this could be tamoxifen resistance alterations that have been acquired by the tumor. Here, we use the "drug view" provided by our web service to evaluate if tamoxifen is a suitable option for the treatment of this tumor (cf. Figure 107).
A closer inspection of molecular markers associated with tamoxifen shows several alterations in key molecules that might impede or reduce the efficacy of the drug. In particular, we observe a frameshift mutation in CYP2D6, a member of the cytochrome P450 family. This enzyme is required to transform tamoxifen into its active form [182]. The frameshift mutation is likely to affect this process and could contribute to potential resistance. Additionally, we also identified a frameshift variant in one of the molecular targets of the drug, namely ESR2, which can also drastically reduce affinity to this receptor. Moreover, we also see an elevated activity of the MAPK signaling pathway and upregulation of HER2/neu (ERBB2). Both alterations could potentially cause a reduced efficacy or even resistance against endocrine therapy by contributing to ligand-independent activation of the estrogen receptor through ERK [439].
In summary, this analysis showed that the investigated tumor carries several alterations that might reduce or impede the function of tamoxifen, which suggests that this drug should not used as a treatment.

C.2.5 *Discussion and conclusion*

In this chapter, we presented two web services for cancer treatment stratification, DrugTargetInspector and ClinOmicsTrail^{bc}. Both web services are designed to assist clinicians with the evaluation of different treatment strategies based on the molecular characteristics of a tumor. For this purpose, our tools offer powerful methods for analyzing high-throughput data sets that can help identify molecular traits of a tumor that are relevant for therapy selection. All results generated by our web services are provided as intuitive and interactive visualizations that facilitate their interpretation. This enables clinicians to evaluate different treatment strategies, to check if a particular drug might be a suitable candidate, or if a drugs mechanisms of action could be impaired.

Although DrugTargetInspector and ClinOmicsTrail^{bc} are currently in a "proof-of-principle" stage, which means they can only be used in a research setting, they already provide powerful functionality that complements existing clinical support tools and that has the potential to improve the treatment stratification process in the future. However, there are still extensions that could further improve their functionality.

For example, in their current form, both tools can assist clinicians to select the targeted drug that seems best suited for the given tumor sample. However, a common problem in cancer therapies that are based on one drug is that eventually patients can develop a resistance to this particular drug, which can even cause a therapy to be ineffective [155]. In order to overcome this problem, several studies suggest using drug combinations instead [107, 155, 363]. Hence, an interesting extension of our web service would be to not only evaluate individual drugs, but also drug combinations that are tailored to a specific patient.

Furthermore, both web services could also be extended with other databases that provide information about disease biomarkers, drug targets, or pharmacogenomic effects, e.g., CIViC [191] or OncoKB [88].

Nevertheless, in conclusion, our web services can already assist clinicians and molecular tumor boards in creating treatment strategies that are specifically tailored to a specific tumor and, hence, constitute valuable tools in the treatment decision-making process for cancer patients.

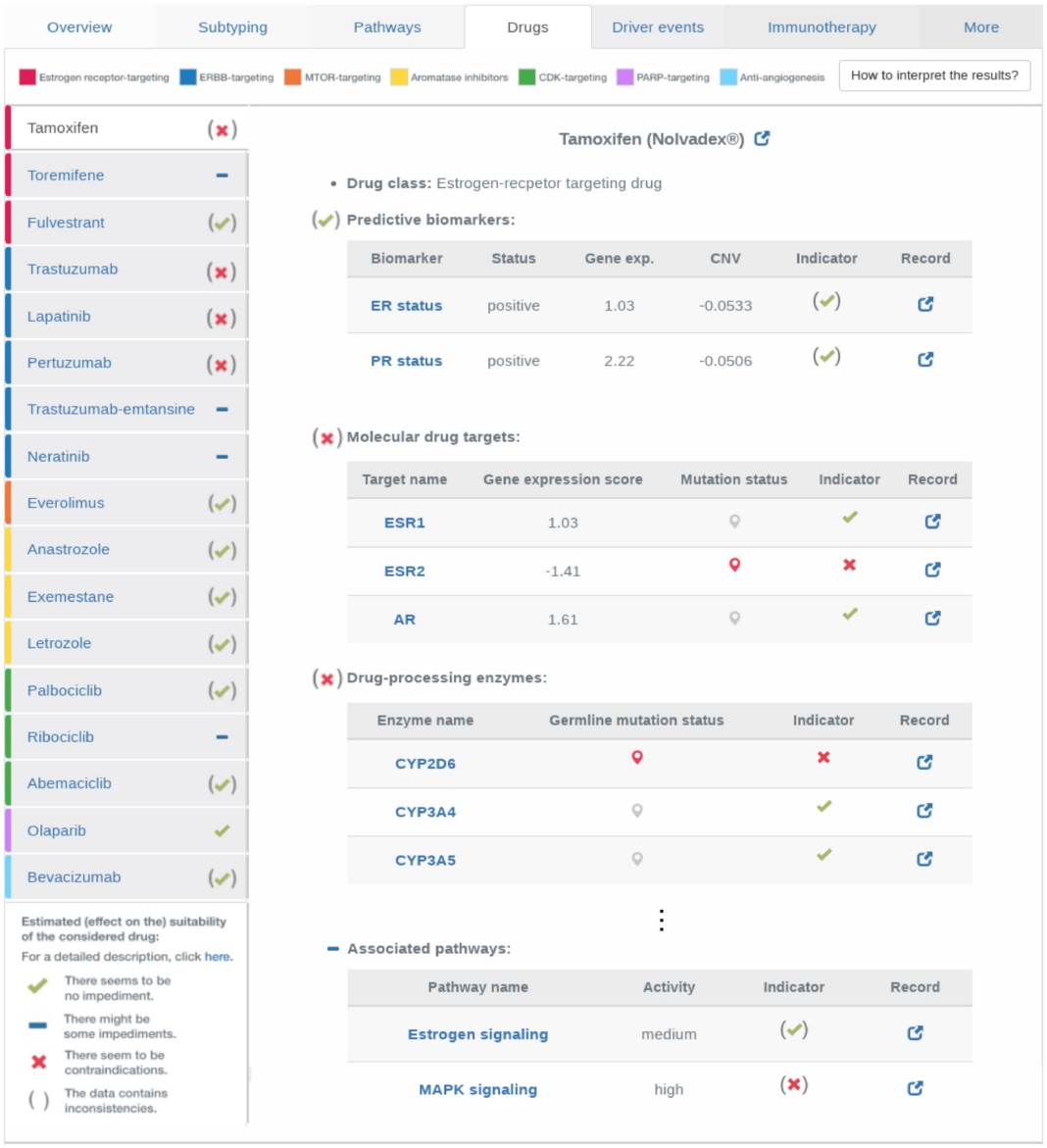


Figure 107: Assessment of standard-of-care breast cancer drugs for TCGA sample "TCGA – BH – A0DT". For each drug, our web server evaluates if predictive biomarkers, molecular drug targets, drug-processing enzymes, resistance-promoting factors, or associated signaling pathways are affected by any (epi)genetic alterations that can influence the efficacy of a drug. Indicator signs show if the different features might have a positive or negative influence on the drug's mode of action. This figure was retrieved from the original ClinOmicsTrail^{bc} publication [475].

D

PROOFS

D.1 SHIFTED EUCLIDEAN DISTANCE

In Section 5.5, we introduced an adapted version of the Euclidean distance called shifted Euclidean distance. For two samples p and q it is defined as follows:

$$d_{shift}(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - s - p_i)^2}, \tag{141}$$

where the optimal value for s is defined as:

$$s = \frac{1}{n} \sum_{i=1}^{n} (q_i - p_i) \tag{142}$$

Here, we provide the proof that s is the optimal value.

Since the square root is a monotonic function, it has no effect on the minimal value and we can omit it to find the optimal value for s. To find the best s, we first determine the critical value of our function:

$$0 = \frac{d}{ds} \sum_{i=1}^{n} (q_i - s - p_i)^2 \tag{143}$$

$$\Leftrightarrow \quad 0 = -2 \sum_{i=1}^{n} (q_i - s - p_i) \tag{144}$$

$$\Leftrightarrow \quad 0 = -2 \sum_{i=1}^{n} (q_i - p_i) + 2ns \tag{145}$$

$$\Leftrightarrow \quad -2ns = -2 \sum_{i=1}^{n} (q_i - p_i) \tag{146}$$

$$\Leftrightarrow \quad s = \frac{1}{n} \sum_{i=1}^{n} (q_i - p_i) \tag{147}$$

Since the second derivative ($\frac{d^2}{ds^2} = 2n$) is always positive, this is the minimum of $d(p, q)$. □

E

ANALYZED DATA SETS, PROCESSING STEPS, AND PARAMETERS

In this chapter, we describe the data sets that are analyzed throughout this thesis as well as the processing steps for all analyses and associated parameters. For some analyses, this chapter also contains additional results and supplementary information that could not be presented, but are referred to, in the main manuscript.

E.1 HEPATOCELLULAR CARCINOMA (HCC) - GSE64041

To demonstrate the capabilities of the GeneTrail standard enrichment analysis functionality, we analyze microarray (Affymetrix Human Gene ST 1.0 arrays) data sets of hepatocellular carcinoma (HCC) patients, control tissue samples, and healthy controls provided by Makowska et al. [327]. This data set contains gene expression measurements of 125 samples from 60 HCC patients and 5 healthy donors. From each patient, two samples were obtained: a tumor biopsy and a paired non-tumor liver biopsy.

E.1.1 *Preprocessing steps*

We downloaded normalized gene expression measurements from the NCBI GEO web server (Accession: GSE64041). To this end, we used the GEO binding integrated in GeneTrail. For normalization the authors used RMA algorithm [239] implementation of oligo package [84]. After the download, the used identifiers were validated and mapped to "Official Gene Symbols" using the mapping strategy outlined in Section 4.3.2. Duplicated entries were replaced by the median value.

E.1.2 *Enrichment analyses*

We used this data set to create the example visualizations of the GeneTrail standard enrichment analysis framework (cf. Section 3.6.1).
For Figure 53, we first compared the gene expression measurements of the tumor biopsis against their paired non-tumor liver samples using the Shrinkage t-test (cf. Section 3.4.3.2). The resulting score list was then used to conduct a GSEA analysis (cf. Section 3.6.3.1). The complete list of parameters can be found in Table 9.

Parameter	Value
Method used for group comparison	Independent shrinkage t-test
Transformation	no transformation
Enrichment analysis method	Kolmogorov-Smirnov test (unweighted GSEA)
Order in which the score list is processed	decreasing
Minimum category size	2
Maximum category size	700
Strategy to calculate p-values	row-wise (exact)
Method used for multiple testing correction	Benjamini-Yekutieli

Table 9: Parameters used for group comparison and subsequent enrichment analysis of the hepatocellular carcinoma (HCC) - GSE64041.

For Figure 52, we first compared the gene expression measurements of three individual tumor biopsis against their paired non-tumor liver sample using $\log_2$-fold-changes. The resulting score list was then used to conduct a GSEA analysis (cf. Section 3.6.3.1). The complete list of parameters can be found in Table 10.

Parameter	Value
Method used for group comparison	$\log_2$ fold-change
Transformation	no transformation
Enrichment analysis method	Kolmogorov-Smirnov test (unweighted GSEA)
Order in which the score list is processed	decreasing
Minimum category size	2
Maximum category size	700
Strategy to calculate p-values	row-wise (exact)
Method used for multiple testing correction	Benjamini-Yekutieli

Table 10: Parameters used to create the result in Figure 52.

E.1.3 *Network analysis*

We used this data set to create the example visualizations of the GeneTrail network analysis framework (cf. Section 3.6.1). To this end, we first compared the gene expression measurements of the tumor biopsis against their paired non-tumor liver samples using the Shrinkage t-test (cf. Section 3.4.3.2). The resulting score list was then used to conduct a subgraph analysis (cf. Section 3.7.1). The complete list of parameters can be found in Table 11.

Parameter	**Value**
Method used for group comparison	Independent shrinkage t-test
Transformation	no transformation
Network analysis method	Subgraph ILP
Subgraph size	10
Scoring mode	absolute value
Node mapping (Complexes)	Minimum
Node mapping (Family)	Minimum

Table 11: Parameters used for group comparison and subsequent network analysis of the hepatocellular carcinoma (HCC) - GSE64041.

E.2 COMPARISON OF ER+ AND ER- BREAST CANCER CELL LINES

Author contributions

The processing steps described in this section are based on our publication 'REGGAE: a novel approach for the identification of key transcriptional regulators' [531].

In this section, we describe the breast cancer data set that was used to compare our REGGAE algorithm to alternative methods in Section 7.2.1. This analysis was part our our publication 'REGGAE: a novel approach for the identification of key transcriptional regulators [531]'. All information in the following paragraphs, was adopted from the respective manuscript or supplementary material.
The gene expression profiles of the breast cancer cell lines were originally published by Heiser et al. [210]. In total, it contains microarray measurements of 46 breast cancer cell lines (Affymetrix GeneChip Human Gene 1.0 ST exon). The ER status of each cell lines was extracted from a study by Neve et al. [385]. Based on the ER status, the cell lines can be categorized into five distinct groups, which are shown in Table 12. For all analyses described in Section 7.2.1, we used all cell lines assigned to Group 1 (ER+) and Group 2 (ER-).

E.2.1 *General processing*

We downloaded the quantile normalized and logarithmized (log2) data set from ArrayExpress (Accession: E-MTAB-181). After the download, the used identifiers were validated and mapped to "Official Gene Symbols" using the mapping strategy outlined in Section 4.3.2. Duplicated entries were replaced by the median value.

E.2.2 *Group comparison and feature selection*

The group comparison (ER+ vs. ER-) was conducted using the Shrinkage t-test (cf. Section 3.4.3.2). Based on the resulting list of differentially expressed genes, we create five test sets for the subsequent analysis: all significantly upregulated genes in ER+ cells ($P < 0.01$) and four sets the most upregulated genes in ER+ cells (250, 500, 750, and 1,000).

Group	ER status	Samples
Group 1	Estrogen-receptor positive (ER+):	600MPE, BT474, BT483, CAMA1, HCC1428, LY2, MCF7, MDAMB134VI, MDAMB175VII, MDAMB361, MDAMB415, T47D, UACC812, ZR751, ZR7530, ZR75B
Group 2	Estrogen-receptor negative (ER-):	AU565, BT20, BT549, HCC38, HCC70, HCC202, HCC1143, HCC1187, HCC1937, HCC1954, HCC2185, HCC3153, HS578T, MCF10A, MCF12A, MDAMB157, MDAMB231, MDAMB453, SKBR3, SUM225CWN, SUM1315MO2
Group 3	Presumably estrogen-receptor positive (ER[+])	SUM52PE
Group 4	Presumably estrogen-receptor negative (ER[-])	SUM149PT, SUM159PT
Group 5	No information available (NA):	184B5, HCC1395, HCC1419, HCC1806, MCF10F, SUM185PE

Table 12: The different sample groups of the data set.

E.2.3 *Regulator effect analyses*

For each method, we carried out an analysis for every test set individually using our entire collection of human RTIs (Version 2, RegulatorTrail). We then aggregated, the five result lists of each method. To this end, we selected the maximum of all p-values and the sum of all ranks. In the last step, the final p-values were FDR adjusted (Benjamini-Yekutieli). In the following paragraphs, we summarize the remaining parameters for each method individually.

Parameter	**Value**
Implementation	RegulatorTrail
P-value strategy	Permutation test with pseudo-count
Number of permutations	1,000,000

Table 13: Parameters used for the correlation set analysis (CSA).

Parameter	**Value**
Implementation	RegulatorTrail
Order of test set	decreasingly with respect to signed t-score
Order of each regulator list	decreasingly with respect to the absolute association score
Association measure	Pearson's correlation coefficient
Enrichment method	Wilcoxon rank-sum test
Number of bootstrap replications	1,000

Table 14: Parameters used for the REGGAE analysis.

Parameter	**Value**
Implementation	RegulatorTrail
Association measure	Pearson's correlation coefficient

Table 15: Parameters used for the RIF1 and RIF2 analyses.

Parameter	**Value**
Implementation	RegulatorTrail
Reference set	All possible target genes in our RTI database

Table 16: Parameters used for the TFactS and TED analyses.

Parameter	**Value**
Implementation	custom Python script

Table 17: Parameters used for the TSS analyses.

Parameter	Value
Implementation	Prototype implementation provided on the authors' web site (http://web.tecnico.ulisboa.pt/aplf/code/tfrank/).

Table 18: Parameters used for the TFRank analyses.

Method	Runtime (s)
CSA[A]	450.27 (±78.76)
REGGAE[B]	174.98 (±1.69)
REGGAE (without bootstrapping)	23.40 (±60.36)
RIF1	23.60 (±0.28)
RIF2	23.85 (±0.10)
TSS	14.86 (±0.63)
TED	658.20 (±29.80)
TFactS	42.37 (±0.23)
TFRank	116.74 (±4.22)

Table 19: Runtime comparison of the different methods. Individual runtimes were obtained on an Intel Core i7-3770 processor using a test set of size 250 and the entire collection of RTIs. (A) CSA analysis for 1,000,000 random permutations. (B) REGGAE analysis with 1,000 bootstrap replications. This Table was adapted from [531].

E.3 CD14 MONOCYTES FROM PERIPHERAL BLOOD OF COVID-19 PATIENTS

Author contributions

The processing steps described in this section are based on our publication "GeneTrail: A Framework for the Analysis of High-Throughput Profiles" [176].

To demonstrate the capabilities of the GeneTrail single-cell workflow, we analyzed a single-cell RNA-seq data set of CD14 monocytes from peripheral blood of COVID-19 patients and healthy controls. The data set contains gene expression profiles of 10,339 cells from blood samples of seven hospitalized patients with COVID-19 and six healthy controls. Of the seven hospitalized patients four were diagnosed with with acute respiratory distress syndrome (ARDS) and required mechanical ventilation.
In this section, we describe the processing steps and parameters used to generate the single-cell enrichment analysis results described in Section 5.6.1. For all analysis we used Version V3.2 of GeneTrail and Version V3 of our database.

E.3.1 *Data download and preprocessing steps*

We downloaded the prefiltered count matrix and associated metadata for each cell from the COVID-19 Cell Atlas [25]("Peripheral Blood MononuclearCells (PBMCs)"). For our analysis, we only consider the subset of CD14 monocytes, which consists of 10,339 cells. The data set was then processed as follows. First, we mapped the provided identifiers to Official gene symbols.

Parameter	**Value**
Method to remove duplicates	median

Table 20: Method used to remove duplicates after id mapping.

E.3.1.1 *Quality control*

We then applied several filter to remove cells with insufficient quality from the data set.

Parameter	**Value**
Minimum number of UMIs	500
Minimum number of expressed genes	500
duplicateMethod	median

Table 21: Parameters used to filter the single-cell data set.

E.3.1.2 *Normalization*

For cells that pass the quality filters, we normalized the expression of all genes.

Parameter	**Value**
Normalization method	log2(RPM+1)

Table 22: Parameters used to normalize the single-cell data set.

E.3.1.3 *Feature selection*

From the normalized expression matrix, we then selected the most expressed genes for each cell.

Parameter	**Value**
Selection method	The X most highly expressed genes
Number of selected genes per cell	500

Table 23: Parameters used to normalize the single-cell data set.

E.3.2 *Enrichment analysis*

After quality control and normalization, we conducted a over-representation analysis for each remaining cell (cf. Section 5.4.2) and subsequently a enrichment-based group comparison (cf. Section 5.4.2.3).

Parameter	**Value**
Test statistic	Hypergeometric test
Method for multiple testing correction	Benjamini-Yekutieli
Minimum number of category members	0
Maximum number of category members	700
Null hypothesis	upper-tailed
Significance level	0.05

Table 24: Parameters used for the single-cell enrichment analysis.

Parameter	**Value**
Test statistic	χ^2-test
Significance level	0.05

Table 25: Parameters used for the enrichment-based group comparison.

E.3.3 *Dimension reduction*

For the visualization of the results, we calculated UMAP coordinates using the Seurat R package [204].

Parameter	**Value**
Most variable genes	2000

Table 26: Parameters used for dimension reduction.

E.4 TIME RESOLVED EXPRESSION PROFILES OF EARLY T CELL ACTIVATION

Author contributions

The processing steps described in this section are based on our publication "GeneTrail 3: advanced high-throughput enrichment analysis" [177].

To demonstrate the capabilities of our time-series workflow, we analyzed a data set of time-resolved gene expression profiles of CD4+ cells that were in vitro activated [121]. The data set consists of 3×13 mRNA expression profiles that were measured in two hour intervals, from 0 to 24 hours, after initial activation.
Here, we desribe the processing steps and parameters used to create the results in Section 5.6.2.

E.4.1 *Microarray experiments (GSE136625)*

RNA isolation and microarray experiments (Agilent-039494 SurePrint G3 Human GE v2 8x60K Microarrays, Cat. no. G4851B, Agilent Technologies, Santa Clara, CA) were conducted by Caroline Diener according to manufacturer instructions and are described in our publication "Quantitative and time-resolved miRNA pattern of early human T cell activation" [121]. In order to process the raw expression values, we used the following processing steps. First, we extracted the raw expression values using the Agilent Feature Extraction Software. For background correction and quantile normalization, we used the limma R-package [498]. After the normalization, the expression values were logarithmized (log2). In a final processing step, we aggregated the three replicates of each time point to reduce the variance.

Parameter	**Value**
Background correction	method = normexp offset = 16
Normalization between the arrays	quantile
Aggregation of technical replicates	median

Table 27: Parameters for the limma normalization and aggregation of the replicates.

E.4.2 *Feature selection and clustering*

To analyze the normalized and aggregated expression values, we conducted the following processing steps. First, we applied a filter step to remove features with no expression change in the analyzed time frame. Next, we applied the two-stage clustering approach described in Section 5.5.

Parameter	**Value**
Difference between minimal and maximal time points	2.0

Table 28: Parameters used for feature selection.

Parameter	**Value**
Distance measure	Euclidean distance for gradients
Linkage method	Complete linkage
Threshold for cluster	0.8
Minimum number of genes for each cluster	1

Table 29: Parameters used for the first stage of the clustering.

Parameter	**Value**
Distance measure	Euclidean distance for gradients
Linkage method	Complete linkage
Threshold for super-cluster	0.95

Table 30: Parameters used for the second stage of the clustering.

E.4.3 *Over-representation analysis (ORA)*

For each cluster, we then applied an over-representation analysis to identify associated deregulated biological processes.

Parameter	**Value**
Minimum number of hits	1
Maximum number of hits	700
Method	Hypergeometric test
P-value strategy	upper-tailed
P-value adjustment	Benjamini-Hochberg
Significance level	0.05

Table 31: Parameters used for the over-representation analysis.

E.5 EXPRESSION PROFILES OF B CELL LYMPHOMAS IN Eμ-MYC TRANSGENIC MICE

Author contributions

The processing steps described in this section are based on our publication 'REGGAE: a novel approach for the identification of key transcriptional regulators' and the respective supplemental material [531].

In order to evaluate the performance of REGGAE (cf. Chapter 7) and alternative approaches (cf. Section 3.8) , we compared microarray expression profiles (Affymerix Mouse Genome 430 2.0 array) of 50 B cell lymphoma samples from Eμ-Myc transgenic mice and 10 healthy lymphnode samples from wild-type mice as a control [369]. The microarrays were processed by the authors using the Affymetrix Microarray Suite version 5.0 (MAS 5.0) with default parameters. The 'global scaling' setting was used to normalize the data set with trimmed mean target intensity of 100 for each sample.

E.5.1 *General processing*

We downloaded the normalized data set from the NCBI GEO repository (Accession: GSE7897). After the download, the used identifiers were validated and mapped to "Official Gene Symbols" using the mapping strategy outlined in Section 4.3.2. Duplicated entries were replaced by the median value.

E.5.2 *Group comparison and feature selection*

The group comparison (lymphoma vs. control) was conducted using the Shrinkage t-test (cf. Section 3.4.3.2). From the resulting score list, we the top 250 most up- and down-regulated genes.

E.5.3 *Regulator effect analyses*

We applied each method to both test sets using the entire collection of human RTIs (Version 2, RegulatorTrail). In the last step, the final p-values were FDR adjusted (Benjamini-Yekutieli). In the following paragraphs, we summarize the remaining parameters for each method individually.

Parameter	Value
Implementation	RegulatorTrail
P-value strategy	Permutation test with pseudo-count
Number of permutations	1,000,000

Table 32: Parameters used for the correlation set analysis (CSA).

Parameter	Value
Implementation	RegulatorTrail
Order of test set	decreasingly with respect to absolute t-score
Order of each regulator list	decreasingly with respect to the absolute association score
Association measure	Pearson's correlation coefficient
Enrichment method	Wilcoxon rank-sum test
Number of bootstrap replications	1,000

Table 33: Parameters used for the REGGAE analysis.

Parameter	Value
Implementation	RegulatorTrail
Association measure	Pearson's correlation coefficient

Table 34: Parameters used for the RIF1 and RIF2 analyses.

Parameter	Value
Implementation	RegulatorTrail
Reference set	All possible target genes in our RTI database

Table 35: Parameters used for the TFactS and TED analyses.

Parameter	Value
Implementation	custom Python script

Table 36: Parameters used for the TSS analyses.

Parameter	Value
Implementation	Prototype implementation provided on the authors' web site (http://web.tecnico.ulisboa.pt/aplf/code/tfrank/).

Table 37: Parameters used for the TFRank analyses.

E.5.4 *Top 25 regulators identified by REGGAE*

The most significant results of REGGAE for the top 250 most up-regulated genes can be found in Table 38. It contains various genes that are known tobe. regulated by MYC (cf. Section 7.2.2.1).

Regulator	Targets	Q-value
MYC	243	8.22×10^{-145}
KAT2A	239	8.66×10^{142}
UBTF	239	1.11×10^{130}
HCFC1	239	5.29×10^{-128}
MBD3	228	5.01×10^{-126}
TFAP4	225	1.57×10^{-119}
RAD23B	221	1.61×10^{-116}
SMC3	239	7.60×10^{-112}
DPY30	189	4.81×10^{-110}
E2F1	231	2.88×10^{-108}
SUZ12	171	8.74×10^{-101}
RCOR2	239	7.87×10^{-100}
SMC1A	225	3.33×10^{-98}
GABPA	239	4.60×10^{-98}
SMARCA5	193	1.41×10^{-97}
MAZ	239	8.97×10^{-97}
SMARCA4	231	3.40×10^{-94}
NELFA	227	1.17×10^{-92}
E2F4	239	2.39×10^{-92}
TBP	239	2.95×10^{-90}
CTCF	239	3.94×10^{-89}
RAD21	239	1.10×10^{-88}
GTF2E2	161	1.57×10^{-84}
TRIM28	136	1.38×10^{-82}
MYB	239	3.05×10^{-82}

Table 38: Excerpt of the REGGAE results for the top 250 up-regualted genes in Eμ-Myc transgenic mice compared to wild-type mice.

E.6 KNOCK-OUT OF PLURIPOTENCY FACTORS IN HUMAN ESCS

Author contributions

The processing steps described in this section are based on our publication 'REGGAE: a novel approach for the identification of key transcriptional regulators' [531].

In order to evaluate the performance of REGGAE (cf. Chapter 7) and competing approaches (cf. Section 3.8), we compared perturbed gene expression profiles of human embryonic stem cells (ESCs) with untreated controls from a data set of Wang et al. [561]. In total, the data set contains 69 microarray profiles (Illumina HumanHT-12 V4.0 expression beadchip). The probe intensities were quantile normalized using the beadarray R package (v1.6) [129]. Of the 69 samples, we used 32: 8 with NANOG knock-out, 8 with POU5F1 (OCT4) knock-out, 8 with SOX2 knock-out, and 8 controls. The respective IDs are listed in Table 39.

E.6.1 *General processing*

We downloaded the quantile normalized data set from the NCBI GEO repository (Accession: GSE34921). After the download, the used identifiers were validated and mapped to "Official Gene Symbols" using the mapping strategy outlined in Section 4.3.2. Duplicated entries were replaced by the median value. We then divided the data set into three subsets that were processed individually: NANOG knock-out + Control, POU5F1 knock-out + Control, and SOX2 knock-out + Control.

E.6.2 *Group comparison and feature selection*

The group comparison (knock-out vs. control) was conducted using the Shrinkage t-test (cf. Section 3.4.3.2). From the resulting score list, we then selected the top 250 most up- and down-regulated genes.

E.6.3 *Regulator effect analyses*

We applied each method to all 6 test sets using the entire collection of human RTIs (Version 2, RegulatorTrail). The resulting p-values of all analyses were FDR adjusted (Benjamini-Yekutieli). In the following paragraphs, we summarize the remaining parameters for each method individually.

GSM number	Sample ID	Group
GSM857336	$H1P_1$	Control
GSM857337	$H1P_2$	Control
GSM857338	$H1P_3$	Control
GSM857339	$H1P_4$	Control
GSM857340	$H1P_5$	Control
GSM857341	$H1P_6$	Control
GSM857342	$H1P_7$	Control
GSM857343	$H1P_8$	Control
GSM857344	$shNANOG_1$	NANOG knock-out
GSM857345	$shNANOG_2$	NANOG knock-out
GSM857346	$shNANOG_3$	NANOG knock-out
GSM857347	$shNANOG_4$	NANOG knock-out
GSM857348	$shNANOG_5$	NANOG knock-out
GSM857349	$shNANOG_6$	NANOG knock-out
GSM857350	$shNANOG_7$	NANOG knock-out
GSM857351	$shNANOG_8$	NANOG knock-out
GSM857352	$shOCT4_1$	POU5F1 knock-out
GSM857353	$shOCT4_2$	POU5F1 knock-out
GSM857354	$shOCT4_3$	POU5F1 knock-out
GSM857355	$shOCT4_4$	POU5F1 knock-out
GSM857356	$shOCT4_5$	POU5F1 knock-out
GSM857357	$shOCT4_6$	POU5F1 knock-out
GSM857358	$shOCT4_7$	POU5F1 knock-out
GSM857359	$shOCT4_8$	POU5F1 knock-out
GSM857360	$shSOX2_1$	SOX2 knock-out
GSM857361	$shSOX2_2$	SOX2 knock-out
GSM857362	$shSOX2_3$	SOX2 knock-out
GSM857363	$shSOX2_4$	SOX2 knock-out
GSM857364	$shSOX2_5$	SOX2 knock-out
GSM857365	$shSOX2_6$	SOX2 knock-out
GSM857366	$shSOX2_7$	SOX2 knock-out
GSM857367	$shSOX2_8$	SOX2 knock-out

Table 39: Samples used for the analysis.

Parameter	Value
Implementation	RegulatorTrail
P-value strategy	Permutation test with pseudo-count
Number of permutations	1,000,000

Table 40: Parameters used for the correlation set analysis (CSA).

Parameter	Value
Implementation	RegulatorTrail
Order of test set	decreasingly with respect to absolute t-score
Order of each regulator list	decreasingly with respect to the absolute association score
Association measure	Pearson's correlation coefficient
Enrichment method	Wilcoxon rank-sum test
Number of bootstrap replications	1,000

Table 41: Parameters used for the REGGAE analysis.

Parameter	Value
Implementation	RegulatorTrail
Association measure	Pearson's correlation coefficient

Table 42: Parameters used for the RIF1 and RIF2 analyses.

Parameter	Value
Implementation	RegulatorTrail
Reference set	All possible target genes in our RTI database

Table 43: Parameters used for the TFactS and TED analyses.

Parameter	Value
Implementation	custom Python script

Table 44: Parameters used for the TSS analyses.

Parameter	Value
Implementation	Prototype implementation provided on the authors' web site (http://web.tecnico.ulisboa.pt/aplf/code/tfrank/).

Table 45: Parameters used for the TFRank analyses.

E.7 KEY REGULATORS IN MACROPHAGES

Author contributions

The processing steps described in this section are based on our publication 'RegulatorTrail: a web service for the identification of key transcriptional regulators' [533].

In order to showcase the capabilities of the RegulatorTrail motif workflow, we analyzed paired chromatin accessibility and gene expression data of macrophages extracted from venous blood. This data set is part of the BLUEPRINT project [334].

E.7.1 *General processing*

We downloaded the processed data set from BLUEPRINT data portal (Accession: S001S745 (DNAse-seq) and S001S712 (RNA-seq)). No further processing was conducted. We uploaded both open chromatin regions and gene expression data directly to RegulatorTrail.

E.7.2 *TEPIC analysis*

The parameters for the TEPIC analysis are summarized in Table 46.

Parameter	**Value**
Reference genome	GRCh38
Window size around 5'-TSS	50 kbp

Table 46: Parameters used for TEPIC.

E.7.3 *INVOKE analysis*

We then used the affinity scores predicted by TEPIC to build a linear model with elastic net penalty (INVOKE, cf. Section 8.1.3.2). To train the model, we used a nested cross-validation with grid search for the parameter α, which balances the lasso and ridge penalty terms. The respective parameters are shown in Table 47. The most influential regulators are shown in Table 48.

Parameter	**Value**
outer cross-validation	6-fold
inner cross-validation	6-fold
step size for α	0.1

Table 47: Parameters used for the INVOKE analysis.

HOXA3	0.148367202	[13, 17, 148]
HLTF	0.089891363	[287]
ETV5	0.078639561	[451]
GMEB1	0.045548794	[254]
HOXA5	0.04494131	[187]
NRF1	0.039230003	[54, 309, 608]
PAX2 (paralog of PAX5)	0.038496525	[213]
ETS2	0.034375522	[323, 599]
ELF5	0.033069267	[538]
NFATC1	0.032528012	[602]
KLF4	0.025539592	[304, 480]
NKX2.5	0.025418796	[432]
RAD21	-0.052709626	[69]

Table 48: Influential transcription factors identified by INVOKE. Depicted are all regulators with regression coefficient of at least 0.025 and references to publications discussing the respective regulators in the context of macrophages. This Figure was adapted from [533].

E.8 WILMS' TUMOR STUDY (EXPRESSION PROFILES)

Author contributions

The processing steps described in this section are based on our publication "The role of TCF3 as potential master regulator in blastemal Wilms tumors." [532]. The wet lab experiments were performed by Nicole Ludwig and the data processing was conducted by me.

For our Wilms' tumor study (cf. Chapter 9), we analyzed gene expression profiles from 33 tumor biopsies of patients treated with neoadjuvant chemotherapy according to the SIOP treatment regimen. The therapy consisted of actinomycin-D, vincristine and, in the case of metastases, doxorubicin [532]. An overview of the clinical characteristics of each sample are shown in Table 50.

E.8.1 *Ethics statement*

"Our study was approved by the ethics committee of the Saarland Medical Council (Ethikkommision der Ärztekammer des Saarlandes, No. 136/01; 09/16/2010). Written informed consent was obtained from the parents of all patients." [532]

E.8.2 *Microarray experiments (GSE98334)*

RNA isolation and microarray experiments (Agilent-039494 SurePrint G3 Human GE v2 8x60K Microarrays, Cat. no. G4851B, Agilent Technologies, Santa Clara, CA) were conducted by Nicole Ludwig according to manufacturer instructions and are described in our publication ' The role of TCF3 as potential master regulator in blastemal Wilms tumors' [532]. Here, we describe the data processing steps used to create the results described in Chapter 9.
In order to process the raw expression values, we used the following processing steps. First, we extracted the raw expression values using the Agilent Feature Extraction Software. For background correction and quantile normalization, we used the limma R-package [498]. Finally, we logarithmized (log2) the normalized expression values.

Parameter	**Value**
Background correction	method = normexp offset = 16
Normalization between the arrays	quantile

Table 49: Parameters for the limma normalization.

GEO sample	Histology	Degree of malignity	Age in months	Gender
WT1	Blastemal subtype	high risk	20	M
WT2	Blastemal subtype	high risk	82	F
WT3	Blastemal subtype	high risk	26	F
WT4	Blastemal subtype	high risk	97	F
WT5	Blastemal subtype	high risk	7	F
WT6	Blastemal subtype	high risk	104	F
WT7	Blastemal subtype	high risk	55	F
WT8	Blastemal subtype	high risk	41	M
WT9	Blastemal subtype	high risk	41	M
WT10	Blastemal subtype	high risk	78	F
WT11	Blastemal subtype	high risk	8	M
WT12	Blastemal subtype	high risk	146	F
WT13	Blastemal subtype	high risk	44	F
WT14	Blastemal subtype	high risk	44	F
WT15	Blastemal subtype	high risk	87	F
WT16	Blastemal subtype	high risk	43	F
WT17	Blastemal subtype	high risk	46	M
WT18	Completely necrotic	low risk	101	F
WT19	Completely necrotic	low risk	42	F
WT20	Diffuse anaplasia	high risk	92	F
WT21	Diffuse anaplasia	high risk	92	F
WT22	Epithelial subtype	intermediate risk	6	M
WT23	Focal anaplasia	intermediate risk	22	M
WT24	Regressive subtype	intermediate risk	64	F
WT25	Regressive subtype	intermediate risk	58	M
WT26	Regressive subtype	intermediate risk	60	F
WT27	Regressive subtype	intermediate risk	116	M
WT28	Regressive subtype	intermediate risk	148	F
WT29	Stromal subtype	intermediate risk	41	F
WT30	Triphasic subtype	intermediate risk	28	F
WT31	Triphasic subtype	intermediate risk	51	F
WT32	Triphasic subtype	intermediate risk	35	F
WT33	Triphasic subtype	intermediate risk	27	M

Table 50: Clinical details for all Wilms' tumor samples in our study. This table was obtained from the supplement of [532].

E.8.3 *Group comparison*

We used the Shrinkage t-test (cf. Section 3.4.3.2) to calculate gene expression differences between blastemal and non-blastemal Wilms tumors. The group assignments of the different samples is shown in Table 50.
Based on the resulting t-score, we created ten sorted score lists. Five lists for the most up-regulated genes, i.e., all significantly upregulated genes ($P(T > t) < 0.01$) and the 250, 500, 750, and 1000 genes with the highest t-score, and five lists for the most downregulated genes that were created analogously.

E.8.4 *REGGAE analysis*

For each of the ten score lists, described in the previous section, we conducted a REGGAE analysis. The p-values for the five lists of up-regulated genes were then aggregated using the second order statistic [123] and FDR adjusted [45]. The five lists of downregulated genes were proceessed accordingly.

Parameter	**Value**
Implementation	RegulatorTrail
Order of test set	decreasingly with respect to absolute t-score
Order of each regulator list	decreasingly with respect to the absolute association score
Association measure	Pearson's correlation coefficient
Enrichment method	Wilcoxon rank-sum test
Number of bootstrap replications	1,000

Table 51: Parameters used for the REGGAE analysis.

E.8.5 *Enrichment analyses*

All enrichment analyses in this study were performed using version 2 of the GeneTrail web service [510].

E.8.5.1 *GSEA of regulator complexes*

For the analysis of regulator complexes, we used gene sets provided by EpiFactors [350] and CORUM [454], as well as custom gene sets of regulators with specific epigenetic functiosn, i.e., regulation of chromatin signaling [248], pluripotency states in ESCs [526], and regulators that occupy super-enhancers in ESCs [212]. The resulting cate-

gories were then used to conduct unweighted gene set enrichment analyses for the most influential regulators obtained in our REGGAE analysis.

Parameter	Value
Minimum number of hits	1
Maximum number of hits	700
Method	Unweighted GSEA with exact p-value
P-value strategy	row-wise (exact)
P-value adjustment	Benjamini-Hochberg
Significance level	0.05

Table 52: Parameters used for the unweighted GSEA of regulator complexes in blastemal Wilm's tumors.

E.8.5.2 *ORA of kidney developmental genes*

For the analysis of kidney developmental genes, we used associated gene sets from our data collection 3.2.3 and custom gene sets from literature [296, 442]. The resulting categories were then used to conduct over-representation analyses for (1) the top 100 regulators in our REGGAE analysis and (2) the top 1000 most upregulated genes in the comparison between blastemal and non-blastemal Wilmstumors. .

Parameter	Value
Minimum number of hits	1
Maximum number of hits	700
Method	Hypergeometric test
P-value strategy	row-wise (exact)
P-value adjustment	Benjamini-Hochberg
Significance level	0.05

Table 53: Parameters used for the ORA of kidney developmental genes in blastemal Wilm's tumors.

E.9 WILMS' TUMOR STUDY (HISTONE MARKS)

Author contributions

The processing steps described in this section are based on our publication "The role of TCF3 as potential master regulator in blastemal Wilms tumors" [532]. The wet lab experiments were conducteed by Kathrin Kattler, Jenny Wegert, and Nicole Ludwig. The data processing was conducted by Kathrin Kattler and me.

For our Wilms' tumor study (cf. Chapter 9), we also analyzed histone marks (H3K4me3 + H3K27ac) of two Wilms' tumor cell lines *ws568li* and *ws998M18*. *ws568li* was derived from an originally stromal Wilm's tumor. The creation and the establishment of this cell line has been described by Wegert et al. [567]. *ws998M18* was derived from a blastemal xenograft tumor the characterization of this cell culture has been described in Kehl et al. [532].
Chromatin preparation, Immunoprecipitation, ChIPmentation, library preparation, and sequencing have also been described in Kehl et al. [532]. Here, we only summarize the computational processing steps. The processed data sets can be downloaded from the GEO repository (Accession: GSE98721).

E.9.1 *Data processing*

The raw sequencing reads were processed as follows. First, we removed low quality ends (phred score = 20) and the adapters of all reads. To this end, we used the Trim Galore software (Version 0.3.3). Next, we used the GEM mapper ([331], Version 1.376 beta) to align the trimmed reads to the human reference genome (Versoin hs37d5). We then converted the alignment from SAM to BAM format using Samtools ([300], Version 1.3). Subsequently, we marked PCR duplications using the MarkDuplicate command from Picard tools (Version 1.115). Finally, we called the peaks using the narrow option of MACS2 ([609], Version 2.1.1) using a minimum FDR cutoff for peak detection (Threshold 0.05).

E.9.1.1 *Comparison of histone marks*

For our manuscript, we compared histone modification patterns (H3K4me3 + H3K27ac) of the two described cell cultures and embryonic stem cells from the Roadmap Epigenomics Mapping Consortium (Epigenome E015). To this end, we first analyzed each sample individually and then compared the results. For this purpose, we assessed for all genes annotated in GENCODE (V27) if their respective promoter regions are affected by one, both, or none of the considered histone modifications in each sample.

We then conducted multiple analysis to identify gene sets with similarities or differences between the investigated samples, cf. Section 9.2.1. For the resulting gene lists, we performed over-representation analysis to identify enriched biological categories.

Parameter	**Value**
Minimum number of hits	1
Maximum number of hits	700
Method	Hypergeometric test
P-value strategy	row-wise (exact)
P-value adjustment	Benjamini-Hochberg
Significance level	0.05
Reference set	All genes annotated in GENCODE release 27 of GRCH37

Table 54: Parameters used for the ORA of kidney developmental genes in blastemal Wilm's tumors.

E.9.1.2 *Integrative analysis of epigenomic and transcriptomic data*

In our study, we also compared genes with H3K4me3 and H3K27ac marks in their promoter region, the most highly expressed genes in blastemal tumors, and target genes of TCF3. For the latter, we used ChIP-Seq experiments of mouse ESCs [97, 333] and mapped them to human orthologs. To compare the different gene sets, we conducted a over-representation analysis using GeneTrail.

Parameter	**Value**
Minimum number of hits	1
Maximum number of hits	700
Method	Hypergeometric test
P-value strategy	row-wise (exact)
P-value adjustment	Benjamini-Hochberg
Significance level	0.05
Reference set	All genes annotated in GENCODE release 27 of GRCH37

Table 55: Parameters used for the ORA of kidney developmental genes in blastemal Wilm's tumors.

E.10 COLON ADENOCARCINOMA ANALYSIS (TCGA-AA-3542)

Author contributions

The processing steps described in this section are based on our publication "DrugTargetInspector: An assistance tool for patient treatment stratification" [474].

In order to demonstrated the capabilities of our DrugTargetInspector web service, we analyzed a colon adenocarcinoma sample (TCGA-AA-3542) from The Cancer Genome Atlas (TCGA) [536]. Here, we describe the general processing steps. These have also been described in our DrugTargetInspector publication [474].

E.10.1 *Download*

We used the TCGA data portal to download normalized expression values (level 3) and processed mutation data (level 2) for the investigated colon adenocarcinoma sample (TCGA-AA-3542) and nine control samples:

Control samples
TCGA-A6-2678-11
TCGA-A6-2683-11
TCGA-AA-3514-11
TCGA-AA-3517-11
TCGA-AA-3520-11
TCGA-AA-3522-11
TCGA-AA-3527-11
TCGA-AA-3531-11
TCGA-AA-3534-11

E.10.2 *Data processing*

For the gene expression data, we calculated a Z-score to calculate the differences between the tumor sample and the panel of controls. The mutation data was downloaded and then converted into VCF format (cf., Section F.2.3). We then used the Ensembl Variant Effect Predictor [346] to annotate all genetic variations for the GRCh37 genome assembly. For the remaining processing steps, we used the default parameters of DrugTargetInspector.

E.11 BREAST CANCER ANALYSIS (TCGA-BH-A0DT)

Author contributions

The processing steps described in this section are based on our publication "ClinOmicsTrail^{bc}: a visual analytics tool for breast cancer treatment stratification" [475].

In order to demonstrate the capabilities of our ClinOmicsTrail^{bc} web service, we analyzed a colon adenocarcinoma sample (TCGA – BH – A0DT) from The Cancer Genome Atlas (TCGA) [536]. Here, we describe the general processing steps. These have also been described in our ClinOmicsTrail^{bc} publication [475].

E.11.1 *Clinical details*

Age	41
Cancer type	Breast cancer
Cancer stage	Stage II
TNM stage	T1 / N1 / M0
Hormone receptor status	ER positive PR positive HER2 negative
Cancer subtype (PAM50)	luminal A

E.11.2 *Download*

We used the TCGA data portal to download clinical information, expression measurements, copy number variations, and mutations from the primary tumor as well as corresponding controls.

E.11.3 *Data processing*

We used the default parameters of ClinOmicsTrail^{bc} to conduct all processing steps.

F

SUPPORTED FILE FORMATS

F.1 FILE FORMATS FOR MOLECULAR MEASUREMENTS

Our web services provide a variety of file formats in which molecular measurements can be supplied.

F.1.1 *Plain text files*

Most measurements can be supplied as plain text files in any whitespace separated format (txt, tsv, ...). All of them assume that each row represents a specific feature.

F.1.1.1 *Feature lists*

Feature list can either contain unordered sets of molecular features or a sorted list of features. While the web services can handle the two cases, they cannot distinguish between them and we rely on the user to select the parameters accordingly.

Listing 6: Example of an unordered feature list.

```
1 BRCA1
2 TCF3
3 SUZ12
```

F.1.1.2 *Score lists*

In addition to the feature name, score lists additionally contain a weight per feature that should reflect its importance, such as a score from a group comparison. This allows users to perform a custom scoring, which might not be supported by our framework.

Listing 7: Example of a score list.

```
1 FOS   3.5
2 JUN   4.5
3 NFKB1 2.2
4 ...
```

F.1.1.3 *Feature matrices*

Feature matrices contain different sample measurements for each feature. Hence, matrices are required to contain a header field that con-

tains a unique identifier for each samples. An example is shown in Listing 5.

Listing 8: Example of a feature matrix with measurements for three genes in the samples.

```
Sample1   Sample2   Control1
CXCL2   0.0   0.1  3.4
IFNG    4.0   4.1  3.9
TNF     5.7   6.5  1.2
...
```

F.1.2 *Sparse matrix formats*

Single-cell data sets often have a large number of samples, but are very sparse. Hence, they are often stored in special matrix formats to save storage space and to reduce network traffic.

F.1.2.1 *Matrix market exchange format (.mtx)*

The Matrix market exchange format is a class of coordinate based file formats, where each file consists of four distinct parts:

1. A header line indicated with %%. This line specifies the format data type.
2. One or multiple comment lines indicated with %
3. One line with three integer values that specify the number of rows, the number of columns, and the number of non-zero entries.
4. Multiple data lines that each represent a coordinate in the matrix (row index, column index, value).

Listing 9: The first few lines of a matrix in matrix market exchange format.

```
%%MatrixMarket matrix coordinate integer general
%metadata_json: {"software_version": "cellranger-arc-1.0.0", "
    format_version": 2}
113843 3012 19472856
25 1 1
63 1 2
100 1 1
169 1 1
...
```

The .mtx format does not support annotations such as row and column names of matrix, hence, they are often provided as additional feature lists (cf. Section F.1.1.1).

F.2 FILE FORMATS FOR GENOMIC INTERVALS

Our framework also supports several file formats to store genetic information that are described in the following sections.

F.2.1 *BED format*

The BED format [260] is a whitespace separated text format used to save genomic regions. It consists of two parts the (optional) header and the body. The format of the header is not clearly defined and depending on the application can fulfill different roles, such as general descriptions of the file (indicated by '#' at the beginning of a line) or instructions for a genome viewer (indicated by 'track' or 'browser' at the beginning of a line). In the body, each line represents a genomic region that is represented by three obligatory columns and up to nine optional ones that can be used to provide additional information, e.g., the name of the region, or the orientation on the genome ('+' or '-' strand).

1. Chromosome identifier (e.g, chr1 or 1)
2. Start coordinate of the interval (inclusive, index starts at 0)
3. End coordinate of the interval (non-inclusive)

Listing 10: Example of a BED file.

```
1 chr1    2140321 2141319
2 chr1    2391041 2392195
3 ...
```

F.2.2 *GFF format*

The general feature format (GFF) is a tab-separated column-based file format used to describe feature annotations of DNA, RNA, or protein sequences, such as the position of genes in a genome. Currently, there are several versions of this format that have slightly different specifications. Here, we focus on version 3 [508].
GFF files contain two parts: an optional header and a body section. The header contains several lines with comments of metadata that are indicated with "##". Each line in the body defines one feature in a given sequence. The features are specified by nine mandatory columns that are described in Table 56.

Column	Description
seqid	The name of the sequence in which the feature occurs.
source	Keyword describing the source of the annotation, e.g., the name of a database.
type	The feature type, e.g., "gene".
start	The start position of the feature in the given sequence (starting with index 1).
end	The end position of the feature in the given sequence.
score	A score for the feature.
strand	A flag ("+" or "-") indicating on which strand the feature is located.
phase	A flag ("0", "1", or "2") that indicates the reading frame of coding sequence (CDS) features, i.e., if the first codon starts at the position 0,1, or 2 of the feature.
attributes	A flag ("0", "1", or "2") that indicates the reading frame of coding sequence (CDS) features, i.e., if the first codon starts at the position 0,1, or 2 of the feature.

Table 56: Mandatory columns in a GFF file (Version 3).

F.2.3 *VCF format*

The Variant Call Format (VCF) is a column-based text format that is used to save genetic variations. It consists of three parts, i.e., multiple metadata lines, a header line, and multiple lines in the body. Each metadata line starts with "##" and contains one key-value pair that describe the file format, metadata of the analyzed sample, or processing steps used to create the file, e.g., applied filters.
The header line starts with "#" and contains the names for the columns in the body. Each line in the body represents one genetic variant. Each line contains eight mandatory fields that describe the associated variant and an arbitrary number fields that represent information about samples. The mandatory fields are:

Column	Description
CHROM	The identifier of the chromosome or contig in which the variant was found.
POS	Position of the first base that is affected by the variant.
ID	A list of unique identifiers for the variant (separated by ";"). Missing values are indicated by ".".
REF	The reference allele, i.e, the base or bases that can be found at the given position at the reference sequence.
ALT	A comma-separated list of alternative allele starting at the given position.
QUAL	A quality score for the variant, i.e., the negative logarithm of the probability that the alternative allele is wrong. Missing values are indicated by ".".
FILTER	A semicolon-separated list of filters that failed when processing the variant, or "PASS" if all filters were passed.
INFO	A list of key-value pairs that provide additional information for the variant.

Table 57: Mandatory columns in a VCF file.

F.2.4 *SEG format*

The segmented data or SEG format is a tab-separated file format for genomic regions and associated scores [397]. It is typically used to represent copy number variations in a genome [397].
A SEG file starts with one header line that specifies thee column names. Each subsequent line represents one genomic region and contains up to six entries (cf. Table 58).

Column	Description
ID	The sample or track name.
chrom	Name of the chromosome or contig that contains the region.
loc.start	Start position of the genomic region.
loc.end	End position of the genomic region.
num.mark	Number of bins covered by the genomic region. (optional)
seg.mean	Average value of the segment. The values are typically in a logarithmic scale.

Table 58: Mandatory columns in a SEG file.

F.2.5 *IDAT format*

Intensity Data (or IDAT) format is an proprietary file format used to save methylation data provided by Illumina Methylation Assays [236]. It contains methylation values for all measured CpGs. To parse those values, specialized tools are required. For all our tools, we use illuminaio [497] to read the data files and RnBeads [371] to calculate the average methylation signal in predefined genomic regions.

F.2.6 *GMT format*

The Gene Matrix Transposed (GMT) format is a text-based file format used to store a collection of biological categories [237, 513]. Each line in a GMT file contains multiple columns that define one specific category. The first column in each row contains the name of the category, the second column contains an information field, and all remaining columns contain the biological entities that belong to this category. The information field can either be a description of the category, am identifier, or a hyperlink to the source. In case of GeneTrail, we often provide a JSON field that contains all available information.

Listing 11: Example of a GMT file containg two categories from the KEGG database. In this case, the information field contains the respective KEGG ids.

```
1 mTOR signaling pathway  hsa04150    IRS1    FZD10   TNF     ...
2 Phagosome               hsa04145    SCARB1  CLEC4M  TFRC    ...
```

G

OVERVIEW OF EXTERNAL RESOURCES

G.1 SUPPORTED ORGANISMS

Latin name	**Common name**	**Taxon ID**
Arabidopsis thaliana	Thale cress	3702
Bos taurus	Cow (cattle)	9913
Caenorhabditis elegans	Roundworm	6239
Canis familiaris	Domestic dog	9615
Danio rerio	Zebrafish	7955
Drosophila melanogaster	Fruit fly	7227
Gallus gallus	Chicken	9031
Homo sapiens	Modern human	9606
Mus musculus	House mouse	10090
Pan troglodytes	Common chimpanzee	9598
Rattus norvegicus	Brown rat	10116
Sus scrofa	Domestic pig	9823

Table 59: Overview of supported organisms. This table is adapted from [177].

Identifier	Example
Ensembl transcript	ENST00000264227
NCBI GI RNA	221041195
RefSeq transcript	XM_005259129
Vega RNA	OTTHUMT00000322976

Table 60: Overview of supported identifier types for transcripts. This table is adapted from [177].

Identifier	Example
Ensembl gene	ENSG00000012048
Entrez gene	602
Gene alias	BCL4
Gene symbol (official)	BCL3
NCBI GI gene	148151281
RefSeq gene	NC_000019
UniGene	Hs.31210
Vega gene	OTTHUMG00000151517

Table 61: Overview of supported identifier types for genes. This table is adapted from [177].

Identifier	Example
Ensembl protein	ENSP00000164227
NCBI GI protein	578822694
UniParc	UPI0000D4AF29
UniProtKB AC/ID	P20749
UniRef50	UniRef50_P20749
UniRef90	UniRef90_P20749
UniRef100	UniRef100_P20749
Vega protein	OTTHUMP00000200151

Table 62: Overview of supported identifier types for proteins. This table is adapted from [177].

Identifier	Example
miRBase Accession	MIMAT0004585
miRBase (V14 - V21)	hsa-miR-15b-3p
miRCarta	m-204

Table 63: Overview of supported identifier types for miRNAs. This table is adapted from [177].

Identifier	Example
dbSNP	rs34039386

Table 64: Overview of supported identifier types for SNPs. This table is adapted from [177].

G.3 DATABASES

Omics	Database
mRNA + Protein	ConsensusPathDB [251], CORUM [454], EpiFactors [350], GO [100], HumanCyc [85], HumanProteinAtlas [415], KEGG [394], miRTarBase [228], MSigDB [306], PANTHER [358], Pfam [41], PharmGKB [211], Reactome [142], RefSeq [421], TRANSFAC [342], WikiPathways [256]
miRNA	GO [100], miRTarBase [228], mirPathDB [530]
SNPs	GWAS Catalog [573], PheWAS Catalog [118]

Table 65: Overview of biological categories for the different omics- types.

Database	Predefined	TSS +/- 1000	TSS +/- 5000	TSS +/- 10000	TSS +1000/ -10000
ChEA[286]	x				
ChipAtlas[396]		x	x	x	
ChipBase[592]		x	x	x	x
ENCODE[99]		x	x	x	x
Jaspar[158]		x	x	x	x
Signalink[146]	x				
TRANSFAC[342]	x				x

Table 66: Overview of the RTIs provided by RegulatorTrail. This table is adapted from [533]

BIBLIOGRAPHY

[1] Dhiraj Acharya, GuanQun Liu, and Michaela U Gack. "Dysregulation of type I interferon responses in COVID-19." In: *Nature Reviews Immunology* 20.7 (2020), pp. 397–398.

[2] Marit Ackermann and Korbinian Strimmer. "A general modular framework for gene set enrichment analysis." In: *BMC bioinformatics* 10.1 (2009), pp. 1–20.

[3] American Cancer Society (ACS). *American Cancer Society (ACS) [Website]*. URL: https://www.cancer.org/cancer.html (visited on 04/30/2021).

[4] Jerry M Adams and Suzanne Cory. "The Bcl-2 apoptotic switch in cancer development and therapy." In: *Oncogene* 26.9 (2007), pp. 1324–1337.

[5] Xian Adiconis, Adam L Haber, Sean K Simmons, Ami Levy Moonshine, Zhe Ji, Michele A Busby, Xi Shi, Justin Jacques, Madeline A Lancaster, Jen Q Pan, et al. "Comprehensive comparative analysis of 5´-end RNA-sequencing methods." In: *Nature methods* 15.7 (2018), pp. 505–511.

[6] Ivan Adzhubei, Daniel M Jordan, and Shamil R Sunyaev. "Predicting functional effect of human missense mutations using PolyPhen-2." In: *Current protocols in human genetics* 76.1 (2013), pp. 7–20.

[7] Affymetrix (Thermo Fisher Scientific Inc.) *Genome-Wide Human SNP Array 5.0*. 2022. URL: https://www.affymetrix.com/products_services/arrays/specific/genome_wide/genome_wide_snp_5.affx (visited on 07/17/2022).

[8] Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. "Predicting effective microRNA target sites in mammalian mRNAs." In: *elife* 4 (2015), e05005.

[9] Agilent. *SurePrint G3 Human CGH Microarray 8x60K*. 2022. URL: https://www.agilent.com/en/product/cgh-cgh-snp-microarray-platform/cgh-cgh-snp-microarrays/human-microarrays/sureprint-g3-human-cgh-microarray-8x60k-228417 (visited on 07/17/2022).

[10] Chiara Agrati, Alessandra Sacchi, Veronica Bordoni, Eleonora Cimini, Stefania Notari, Germana Grassi, Rita Casetti, Eleonora Tartaglia, Eleonora Lalle, Alessandra D'Abramo, et al. "Expansion of myeloid-derived suppressor cells in patients with severe coronavirus disease (COVID-19)." In: *Cell Death & Differentiation* 27.11 (2020), pp. 3196–3207.

[11] Aviva Presser Aiden, Miguel N Rivera, Esther Rheinbay, Manching Ku, Erik J Coffman, Thanh T Truong, Sara O Vargas, Eric S Lander, Daniel A Haber, and Bradley E Bernstein. "Wilms tumor chromatin profiles highlight stem cell properties and a renal developmental network." In: *Cell stem cell* 6.6 (2010), pp. 591–602.

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

[13] Hadeel Al Sadoun, Matthew Burgess, Kathryn E Hentges, and Kimberly A Mace. "Enforced expression of Hoxa3 inhibits classical and promotes alternative activation of macrophages in vitro and in vivo." In: *The Journal of Immunology* 197.3 (2016), pp. 872–884.

[14] Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Peter Walter Keith Roberts, et al. "Molecular biology of the cell." In: (2018).

[15] Julia Alles, Tobias Fehlmann, Ulrike Fischer, Christina Backes, Valentina Galata, Marie Minet, Martin Hart, Masood Abu-Halima, Friedrich A Grässer, Hans-Peter Lenhof, et al. "An estimate of the total number of true human miRNAs." In: *Nucleic acids research* 47.7 (2019), pp. 3353–3364.

[16] Roberto Alonso, Francisco Salavert, Francisco Garcia-Garcia, Jose Carbonell-Caballero, Marta Bleda, Luz Garcia-Alonso, Alba Sanchis-Juan, Daniel Perez-Gil, Pablo Marin-Garcia, Ruben Sanchez, et al. "Babelomics 5.0: functional interpretation for new generations of genomic data." In: *Nucleic acids research* 43.W1 (2015), W117–W121.

[17] Salma Alrdahe, Hadeel Al Sadoun, Tanja Torbica, Edward A McKenzie, Frank L Bowling, Andrew JM Boulton, and Kimberly A Mace. "Dysregulation of macrophage development and phenotype in diabetic human macrophages can be rescued by Hoxa3 protein transduction." In: *PloS one* 14.10 (2019), e0223980.

[18] Alicia Amadoz, Patricia Sebastian-Leon, Enrique Vidal, Francisco Salavert, and Joaquin Dopazo. "Using activation status of signaling pathways as mechanism-based biomarkers to predict drug sensitivity." In: *Scientific reports* 5.1 (2015), pp. 1–14.

[19] Renée van Amerongen and Roel Nusse. "Towards an integrated view of Wnt signaling in development." In: *Development* 136.19 (2009), pp. 3205–3214.

[20] Simon Anders and Wolfgang Huber. "Differential expression analysis for sequence count data." In: *Nature Precedings* (2010), pp. 1–1.

[21] Sharon Anderson, Alan T Bankier, Bart G Barrell, Maarten HL de Bruijn, Alan R Coulson, Jacques Drouin, Ian C Eperon, Donald P Nierlich, Bruce A Roe, Frederick Sanger, et al. "Sequence and organization of the human mitochondrial genome." In: *Nature* 290.5806 (1981), pp. 457–465.

[22] Guillermo Arango Duque and Albert Descoteaux. "Macrophage cytokines: involvement in immunity and infectious diseases." In: *Frontiers in immunology* 5 (2014), p. 491.

[23] Prabhu S Arunachalam, Florian Wimmers, Chris Ka Pun Mok, Ranawaka APM Perera, Madeleine Scott, Thomas Hagan, Natalia Sigal, Yupeng Feng, Laurel Bristow, Owen Tak-Yin Tsang, et al. "Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans." In: *Science* 369.6508 (2020), pp. 1210–1220.

[24] Highsoft AS. *HighCharts*. URL: https://www.highcharts.com/ (visited on 03/11/2021).

[25] Wellcome Trust Human Cell Atlas. *COVID-19 Cell Atlas*. 2022. URL: https://www.covid19cellatlas.org/ (visited on 01/08/2022).

[26] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. "Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III." In: *The Journal of experimental medicine* 79.2 (1944), pp. 137–158.

[27] Joshua E Babiarz, J Graham Ruby, Yangming Wang, David P Bartel, and Robert Blelloch. "Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs." In: *Genes & development* 22.20 (2008), pp. 2773–2785.

[28] Christina Backes, Tobias Fehlmann, Fabian Kern, **Tim Kehl**, Hans-Peter Lenhof, Eckart Meese, and Andreas Keller. "miRCarta: a central repository for collecting miRNA candidates." In: *Nucleic acids research* 46.D1 (2018), pp. D160–D167.

[29] Christina Backes, Andreas Keller, Jan Kuentzer, Benny Kneissl, Nicole Comtesse, Yasser A Elnakady, Rolf Müller, Eckart Meese, and Hans-Peter Lenhof. "GeneTrail—advanced gene set enrichment analysis." In: *Nucleic acids research* 35.suppl_2 (2007), W186–W192.

[30] Christina Backes, Qurratulain T Khaleeq, Eckart Meese, and Andreas Keller. "miEAA: microRNA enrichment analysis and annotation." In: *Nucleic acids research* 44.W1 (2016), W110–W116.

[31] Christina Backes, Eckart Meese, Hans-Peter Lenhof, and Andreas Keller. "A dictionary on microRNAs and their putative target pathways." In: *Nucleic acids research* 38.13 (2010), pp. 4476–4486.

[32] Christina Backes, Alexander Rurainski, Gunnar W Klau, Oliver Müller, Daniel Stöckel, Andreas Gerasch, Jan Küntzer, Daniela Maisel, Nicole Ludwig, Matthias Hein, et al. "An integer linear programming approach for finding deregulated subgraphs in regulatory networks." In: *Nucleic acids research* 40.6 (2012), e43–e43.

[33] Christina Backes, **Tim Kehl**, Daniel Stöckel, Tobias Fehlmann, Lara Schneider, Eckart Meese, Hans-Peter Lenhof, and Andreas Keller. "miRPathDB: a new dictionary on microRNAs and target pathways." In: *Nucleic acids research* (2016), gkw926.

[34] Soo Youn Bae, Sangmin Kim, Jun Ho Lee, Hyun-chul Lee, Se Kyung Lee, Won Ho Kil, Seok Won Kim, Jeong Eon Lee, and Seok Jin Nam. "Poor prognosis of single hormone receptor-positive breast cancer: similar outcome as triple-negative breast cancer." In: *BMC cancer* 15.1 (2015), pp. 1–9.

[35] Yu-Long Bai, Melody Baddoo, Erik K Flemington, Hani N Nakhoul, and Yao-Zhong Liu. "Screen technical noise in single cell RNA sequencing data." In: *Genomics* 112.1 (2020), pp. 346–355.

[36] Timothy L Bailey. "DREME: motif discovery in transcription factor ChIP-seq data." In: *Bioinformatics* 27.12 (2011), pp. 1653–1659.

[37] Andrew J Bannister and Tony Kouzarides. "Regulation of chromatin by histone modifications." In: *Cell research* 21.3 (2011), pp. 381–395.

[38] Nehla Banu, Sandeep Surendra Panikar, Lizbeth Riera Leal, and Annie Riera Leal. "Protective role of ACE2 and its downregulation in SARS-CoV-2 infection leading to macrophage activation syndrome: therapeutic implications." In: *Life sciences* 256 (2020), p. 117905.

[39] Irena Barbulovic-Nad, Michael Lucente, Yu Sun, Mingjun Zhang, Aaron R Wheeler, and Markus Bussmann. "Bio-microarray fabrication techniques—a review." In: *Critical reviews in biotechnology* 26.4 (2006), pp. 237–259.

[40] David P Bartel. "Metazoan micrornas." In: *Cell* 173.1 (2018), pp. 20–51.

[41] Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik LL Sonnhammer, et al. "The Pfam protein families database." In: *Nucleic acids research* 32.suppl_1 (2004), pp. D138–D141.

[42] Carl Benda. "Ueber die spermatogenese der vertebraten und höherer evertebraten, II. Theil: Die histiogenese der spermien." In: *Arch. Anat. Physiol* 73 (1898), pp. 393–398.

[43] Martin Bengtsson, Anders Ståhlberg, Patrik Rorsman, and Mikael Kubista. "Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels." In: *Genome research* 15.10 (2005), pp. 1388–1392.

[44] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[45] Yoav Benjamini and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency." In: *Annals of statistics* (2001), pp. 1165–1188.

[46] Otto G Berg and Peter H von Hippel. "Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters." In: *Journal of molecular biology* 193.4 (1987), pp. 723–743.

[47] Gabriele Bergers and Laura E Benjamin. "Tumorigenesis and the angiogenic switch." In: *Nature reviews cancer* 3.6 (2003), pp. 401–410.

[48] Tim Berners-Lee. *Universal resource identifiers in WWW*. 1994.

[49] Tim Berners-Lee, Robert Cailliau, Ari Luotonen, Henrik Frystyk Nielsen, and Arthur Secret. "The world-wide web." In: *Communications of the ACM* 37.8 (1994), pp. 76–82.

[50] Tim Berners-Lee, Roy Fielding, Larry Masinter, et al. *Uniform resource identifiers (URI): Generic syntax*. 1998.

[51] Tim Berners-Lee, Larry Masinter, Mark McCahill, et al. "Uniform resource locators (URL)." In: (1994).

[52] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, et al. "The NIH roadmap epigenomics mapping consortium." In: *Nature biotechnology* 28.10 (2010), pp. 1045–1048.

[53] Doron Betel, Anjali Koppal, Phaedra Agius, Chris Sander, and Christina Leslie. "Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites." In: *Genome biology* 11.8 (2010), pp. 1–14.

[54] TA Beyer, Ulrich Auf dem Keller, S Braun, M Schäfer, and S Werner. "Roles and mechanisms of action of the Nrf2 transcription factor in skin morphogenesis, wound repair and skin cancer." In: *Cell death and differentiation* 14.7 (2007), p. 1250.

[55] BGI. *BGISEQ*. 2020. URL: https://www.bgi.com/dev/wp-content/uploads/sites/19/2017/04/BGISEQ-500-WGS-Technical-Note-Flyer_20180223A4\%E5\%87\%BA\%E8\%A1\%80.pdf (visited on 12/26/2020).

[56] Anindya Bhattacharya and Yan Cui. "miR2GO: comparative functional analysis for microRNAs." In: *Bioinformatics* 31.14 (2015), pp. 2403–2405.

[57] Anindya Bhattacharya, Jesse D Ziebarth, and Yan Cui. "PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways." In: *Nucleic acids research* 42.D1 (2014), pp. D86–D91.

[58] Nirmala Bhoo-Pathy, Cheng-Har Yip, Mikael Hartman, Nakul Saxena, Nur Aishah Taib, Gwo-Fuang Ho, Lai-Meng Looi, Awang M Bulgiba, Yolanda van der Graaf, and Helena M Verkooijen. "Adjuvant! Online is overoptimistic in predicting survival of Asian breast cancer patients." In: *European Journal of Cancer* 48.7 (2012), pp. 982–989.

[59] Eva Bianconi, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, Maria Chiara Pelleri, Simone Tassani, Francesco Piva, et al. "An estimation of the number of cells in the human body." In: *Annals of human biology* 40.6 (2013), pp. 463–471.

[60] Mithun Biswas, Karine Voltz, Jeremy C Smith, and Jörg Langowski. "Role of histone tails in structural stability of the nucleosome." In: *PLoS computational biology* 7.12 (2011), e1002279.

[61] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. "Fast unfolding of communities in large networks." In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[62] Christoph Bock, Matthias Farlik, and Nathan C Sheffield. "Multi-omics of single cells: strategies and applications." In: *Trends in biotechnology* 34.8 (2016), pp. 605–608.

[63] Carlo Bonferroni. "Teoria statistica delle classi e calcolo delle probabilita." In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8 (1936), pp. 3–62.

[64] Carlo E Bonferroni. "Il calcolo delle assicurazioni su gruppi di teste." In: *Studi in onore del professore salvatore ortu carboni* (1935), pp. 13–60.

[65] WA Bonner, HR Hulett, RG Sweet, and LA Herzenberg. "Fluorescence activated cell sorting." In: *Review of Scientific Instruments* 43.3 (1972), pp. 404–409.

[66] Bert Bos, Tantek Çelik, Ian Hickson, and Håkon Wium Lie. "Cascading style sheets level 2 revision 1 (css 2.1) specification." In: *W3C working draft, W3C, June* (2005).

[67] Mike Bostock. *D3*. URL: https://d3js.org/ (visited on 03/11/2021).

[68] Emmanuel Boutet, Damien Lieberherr, Michael Tognolli, Michel Schneider, and Amos Bairoch. "Uniprotkb/swiss-prot." In: *Plant bioinformatics*. Springer, 2007, pp. 89–112.

[69] Sarion R Bowers, Fabio Mirabella, Fernando J Calero-Nieto, Stephanie Valeaux, Suzana Hadjur, Euan W Baxter, Matthias Merkenschlager, and Peter N Cockerill. "A conserved insulator that recruits CTCF and cohesin exists between the closely related but divergently regulated interleukin-3 and granulocyte-macrophage colony-stimulating factor genes." In: *Molecular and cellular biology* 29.7 (2009), pp. 1682–1693.

[70] Linda M Boxer and Chi V Dang. "Translocations involving c-myc and c-myc function." In: *Oncogene* 20.40 (2001), pp. 5595–5610.

[71] Alan P Boyle, Lingyun Song, Bum-Kyu Lee, Darin London, Damian Keefe, Ewan Birney, Vishwanath R Iyer, Gregory E Crawford, and Terrence S Furey. "High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells." In: *Genome research* 21.3 (2011), pp. 456–464.

[72] Robert Braden et al. "Requirements for internet hosts–communication layers." In: (1989).

[73] Robert Braden et al. "Requirements for Internet hosts-application and support." In: (1989).

[74] Tim Bray et al. "The javascript object notation (json) data interchange format." In: (2014).

[75] Michael Brenowitz, Donald F Senear, and Robert E Kingston. "DNase I footprint analysis of protein-DNA binding." In: *Current protocols in molecular biology* 7.1 (1989), pp. 12–4.

[76] Robert Brown. *Observations on the organs and mode of fecundation in Orchideae and Asclepiadeae*. Taylor, 1833.

[77] Brigitte Bruijns, Roald Tiggelaar, and Han Gardeniers. "Massively parallel sequencing techniques for forensics: a review." In: *Electrophoresis* 39.21 (2018), pp. 2642–2654.

[78] Eric W Brunskill, Bruce J Aronow, Kylie Georgas, Bree Rumballe, M Todd Valerius, Jeremy Aronow, Vivek Kaimal, Anil G Jegga, Jing Yu, Sean Grimmond, et al. "Atlas of Gene Expression in the Developing Kidney at Microanatomic Resolution (vol 15, pg 781, 2008)." In: *DEVELOPMENTAL CELL* 16.3 (2009), pp. 482–482.

[79] Philipp Bucher. "Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences." In: *Journal of molecular biology* 212.4 (1990), pp. 563–578.

[80] David M Budden, Daniel G Hurley, Joseph Cursons, John F Markham, Melissa J Davis, and Edmund J Crampin. "Predicting expression: the complementary power of histone modification and transcription factor binding data." In: *Epigenetics & chromatin* 7.1 (2014), pp. 1–12.

[81] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. "Single-cell chromatin accessibility reveals principles of regulatory variation." In: *Nature* 523.7561 (2015), pp. 486–490.

[82] Ken M Cadigan and Marian L Waterman. "TCF/LEFs and Wnt signaling in the nucleus." In: *Cold Spring Harbor perspectives in biology* 4.11 (2012), a007906.

[83] Vincent C Calhoun, Angelike Stathopoulos, and Michael Levine. "Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex." In: *Proceedings of the National Academy of Sciences* 99.14 (2002), pp. 9243–9247.

[84] Benilton S Carvalho and Rafael A Irizarry. "A framework for oligonucleotide microarray preprocessing." In: *Bioinformatics* 26.19 (2010), pp. 2363–2367.

[85] Ron Caspi, Richard Billington, Hartmut Foerster, Carol A Fulcher, Ingrid Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A Mueller, Quang Ong, et al. "Biocyc: Online resource for genome and metabolic pathway analysis." In: *The FASEB Journal* 30 (2016), lb192–lb192.

[86] JM Cavaillon. "Cytokines and macrophages." In: *Biomedicine & pharmacotherapy* 48.10 (1994), pp. 445–453.

[87] Vinton G Cerf and Edward Cain. "The DoD internet architecture model." In: *Computer Networks (1976)* 7.5 (1983), pp. 307–318.

[88] Debyani Chakravarty, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E Rudolph, Rona Yaeger, Tara Soumerai, Moriah H Nissan, et al. "OncoKB: a precision oncology knowledge base." In: *JCO precision oncology* 1 (2017), pp. 1–16.

[89] Nicole M Chapman and Hongbo Chi. "Hallmarks of T-cell Exit from Quiescence." In: *Cancer immunology research* 6.5 (2018), pp. 502–508.

[90] Jocelyn Charlton, Vesna Pavasovic, and Kathy Pritchard-Jones. "Biomarkers to detect Wilms tumors in pediatric patients: where are we now?" In: *Future Oncology* 11.15 (2015), pp. 2221–2234.

[91] AO Chiromatzo, TYK Oliveira, G Pereira, AY Costa, CAE Montesco, DE Gras, F Yosetake, JB Vilar, M Cervato, PRR Prado, et al. "miRNApath: a database of miRNAs, target genes and metabolic pathways." In: *Genetics and Molecular Research* (2007), pp. 859–865.

[92] Anshika Chowdhary, Venkata Satagopam, and Reinhard Schneider. "Long non-coding RNAs: mechanisms, experimental, and computational approaches in identification, characterization, and their biomarker potential in cancer." In: *Frontiers in Genetics* 12 (2021), p. 770.

[93] Jared M Churko, Gary L Mantalas, Michael P Snyder, and Joseph C Wu. "Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases." In: *Circulation research* 112.12 (2013), pp. 1613–1623.

[94] Cedric R Clapier, Janet Iwasa, Bradley R Cairns, and Craig L Peterson. "Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes." In: *Nature reviews Molecular cell biology* 18.7 (2017), pp. 407–422.

[95] Stephen J Clark, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M Stubbs, Heather J Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C Marioni, et al. "scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells." In: *Nature communications* 9.1 (2018), pp. 1–9.

[96] John P Cogswell, James Ward, Ian A Taylor, Michelle Waters, Yunling Shi, Brian Cannon, Kevin Kelnar, Jon Kemppainen, David Brown, Caifu Chen, et al. "Identification of miRNA changes in Alzheimer's disease brain and CSF yields putative biomarkers and insights into disease pathways." In: *Journal of Alzheimer's disease* 14.1 (2008), pp. 27–41.

[97] Megan F Cole, Sarah E Johnstone, Jamie J Newman, Michael H Kagey, and Richard A Young. "Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells." In: *Genes & development* 22.6 (2008), pp. 746–755.

[98] David Colquhoun. *Lectures on biostatistics: an introduction to statistics with applications in biology and medicine*. David Colquhoun, 1971, pp. 96–97.

[99] ENCODE Project Consortium et al. "The ENCODE (ENCyclopedia of DNA elements) project." In: *Science* 306.5696 (2004), pp. 636–640.

[100] Gene Ontology Consortium. "The Gene Ontology (GO) database and informatics resource." In: *Nucleic acids research* 32.suppl_1 (2004), pp. D258–D261.

[101] Tabula Muris Consortium et al. "Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris." In: *Nature* 562.7727 (2018), pp. 367–372.

[102] Marie Corbin, Aurélien de Reyniès, David S Rickman, Dominique Berrebi, Liliane Boccon-Gibod, Sarah Cohen-Gogo, Monique Fabre, Francis Jaubert, Marine Faussillon, Funda Yilmaz, et al. "WNT/β-catenin pathway activation in Wilms tumors: A unifying mechanism with multiple entries?" In: *Genes, Chromosomes and Cancer* 48.9 (2009), pp. 816–827.

[103] IG Costa, HG Roider, and TG do Rego. "F. d. A. de Carvalho,"Predicting gene expression in t cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models,"" in: *BMC bioinformatics* 12.1 (2011), S29.

[104] Jennifer Couzin-Frankel. *Cancer immunotherapy*. 2013.

[105] Francis HC Crick. "On protein synthesis." In: *Symp Soc Exp Biol*. Vol. 12. 138-63. 1958, p. 8.

[106] Francis HC Crick, Leslie Barnett, Sydney Brenner, and Richard J Watts-Tobin. "General nature of the genetic code for proteins." In: *Nature* 192.4809 (1961), pp. 1227–1232.

[107] Adam S Crystal, Alice T Shaw, Lecia V Sequist, Luc Friboulet, Matthew J Niederst, Elizabeth L Lockerman, Rosa L Frias, Justin F Gainor, Arnaud Amzallag, Patricia Greninger, et al. "Patient-derived models of acquired resistance can identify effective drug combinations for cancer." In: *Science* 346.6216 (2014), pp. 1480–1486.

[108] Leonardo Dagum and Ramesh Menon. "OpenMP: an industry standard API for shared-memory programming." In: *IEEE computational science and engineering* 5.1 (1998), pp. 46–55.

[109] Chi V Dang. "MYC on the path to cancer." In: *Cell* 149.1 (2012), pp. 22–35.

[110] James E Darnell. "Transcription factors as targets for cancer therapy." In: *Nature Reviews Cancer* 2.10 (2002), pp. 740–749.

[111] Andrew M Davidoff. "Wilms tumor." In: *Advances in Pediatrics* 59.1 (2012), pp. 247–267.

[112] Morris H DeGroot and Mark J. Schervish. *Probability and statistics*. Pearson Education, 2012.

[113] Scott M Dehm. "A causal role for ERG in neoplastic transformation of prostate epithelium: Klezovitch O, Risk M, Coleman I, Lucas JM, Null M, True LD, Nelson PS, Vasioukhin V, Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA." In: *Urologic Oncology: Seminars and Original Investigations*. Vol. 26. 6. Elsevier. 2008, p. 688.

[114] Benjamin Dekel, Sally Metsuyanim, Kai M Schmidt-Ott, Edi Fridman, Jasmin Jacob-Hirsch, Amos Simon, Jehonathan Pinthus, Yoram Mor, Jonathan Barasch, Ninette Amariglio, et al. "Multiple imprinted and stemness genes provide a link between normal and tumor progenitor cells of the developing human kidney." In: *Cancer research* 66.12 (2006), pp. 6040–6049.

[115] Giusy Della Gatta, Teresa Palomero, Arianne Perez-Garcia, Alberto Ambesi-Impiombato, Mukesh Bansal, Zachary W Carpenter, Kim De Keersmaecker, Xavier Sole, Luyao Xu, Elisabeth Paietta, et al. "Reverse engineering of TLX oncogenic transcriptional networks identifies RUNX1 as tumor suppressor in T-ALL." In: *Nature medicine* 18.3 (2012), pp. 436–440.

[116] Peter J Delves and Ivan M Roitt. "The immune system." In: *New England journal of medicine* 343.1 (2000), pp. 37–49.

[117] Ahmet M Denli, Bastiaan BJ Tops, Ronald HA Plasterk, René F Ketting, and Gregory J Hannon. "Processing of primary microRNAs by the Microprocessor complex." In: *Nature* 432.7014 (2004), pp. 231–235.

[118] Joshua C Denny, Lisa Bastarache, Marylyn D Ritchie, Robert J Carroll, Raquel Zink, Jonathan D Mosley, Julie R Field, Jill M Pulley, Andrea H Ramirez, Erica Bowton, et al. "Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data." In: *Nature biotechnology* 31.12 (2013), pp. 1102–1111.

[119] SA Denslow and PA Wade. "The human Mi-2/NuRD complex and gene regulation." In: *Oncogene* 26.37 (2007), pp. 5433–5438.

[120] Rafael Hernández-de Diego, Sonia Tarazona, Carlos Martínez-Mira, Leandro Balzano-Nogueira, Pedro Furió-Tarí, Georgios J Pappas Jr, and Ana Conesa. "PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data." In: *Nucleic acids research* 46.W1 (2018), W503–W509.

[121] Caroline Diener, Martin Hart, **Tim Kehl**, Stefanie Rheinheimer, Nicole Ludwig, Lena Krammes, Sarah Pawusch, Kerstin Lenhof, Tanja Tänzer, David Schub, et al. "Quantitative and time-resolved miRNA pattern of early human T cell activation." In: *Nucleic Acids Research* (2020).

[122] Johanna K DiStefano. "The emerging role of long noncoding RNAs in human disease." In: *Disease Gene Identification* (2018), pp. 91–110.

[123] Marcus T Dittrich, Gunnar W Klau, Andreas Rosenwald, Thomas Dandekar, and Tobias Müller. "Identifying functional modules in protein–protein interaction networks: an integrated exact approach." In: *Bioinformatics* 24.13 (2008), pp. i223–i231.

[124] Jeffrey S Dome, Elizabeth J Perlman, and Norbert Graf. "Risk stratification for wilms tumor: current approach and future directions." In: *American Society of Clinical Oncology Educational Book* 34.1 (2014), pp. 215–223.

[125] Bruno Domon and Ruedi Aebersold. "Mass spectrometry and protein analysis." In: *science* 312.5771 (2006), pp. 212–217.

[126] Francisco J Candido Dos Reis, Gordon C Wishart, Ed M Dicks, David Greenberg, Jem Rashbass, Marjanka K Schmidt, Alexandra J van den Broek, Ian O Ellis, Andrew Green, Emad Rakha, et al. "An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation." In: *Breast Cancer Research* 19.1 (2017), pp. 1–13.

[127] Sorin Drăghici, Purvesh Khatri, Rui P Martins, G Charles Ostermeier, and Stephen A Krawetz. "Global functional profiling of gene expression." In: *Genomics* 81.2 (2003), pp. 98–104.

[128] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification and scene analysis*. Vol. 3. Wiley New York, 1973.

[129] Mark J Dunning, Mike L Smith, Matthew E Ritchie, and Simon Tavaré. "beadarray: R classes and methods for Illumina bead-based data." In: *Bioinformatics* 23.16 (2007), pp. 2183–2184.

[130] Lydia Dyck and Kingston HG Mills. "Immune checkpoints and their inhibition in cancer and infectious diseases." In: *European journal of immunology* 47.5 (2017), pp. 765–779.

[131] William S Dynan. "DNase I footprinting as an assay for mammalian gene regulatory proteins." In: *Genetic engineering*. Springer, 1987, pp. 75–87.

[132] Final Draft ECMA. "ECMAScript Language Specification." In: (1999).

[133] Dominic Edelmann, Tamás F Móri, and Gábor J Székely. "On relationships between the Pearson and the distance correlation coefficients." In: *Statistics & Probability Letters* 169 (2021), p. 108960.

[134] Bradley Efron. "Better bootstrap confidence intervals." In: *Journal of the American statistical Association* 82.397 (1987), pp. 171–185.

[135] Bradley Efron. "Bootstrap methods: another look at the jackknife." In: *Breakthroughs in statistics*. Springer, 1992, pp. 569–593.

[136] Bradley Efron, Robert Tibshirani, et al. "On testing the significance of sets of genes." In: *The annals of applied statistics* 1.1 (2007), pp. 107–129.

[137] Hussein I El-Subbagh and Abdullah A Al-Badr. "Cytarabine." In: *Profiles of Drug Substances, Excipients and Related Methodology*. Vol. 34. Elsevier, 2009, pp. 37–113.

[138] AJ Enright, B John, U Gaul, T Tuschl, C Sander, and DS Marks. *MicroRNA targets in Drosophila. genome biol*. 2003.

[139] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, et al. "Mapping and analysis of chromatin state dynamics in nine human cell types." In: *Nature* 473.7345 (2011), pp. 43–49.

[140] Virginia Espina, Michael Heiby, Mariaelena Pierobon, and Lance A Liotta. "Laser capture microdissection technology." In: *Expert review of molecular diagnostics* 7.5 (2007), pp. 647–657.

[141] Ahmed Essaghir, Federica Toffalini, Laurent Knoops, Anders Kallin, Jacques van Helden, and Jean-Baptiste Demoulin. "Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data." In: *Nucleic acids research* 38.11 (2010), e120–e120.

[142] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. "The reactome pathway knowledgebase." In: *Nucleic acids research* 46.D1 (2018), pp. D649–D655.

[143] Yannan Fan, Keith Siklenka, Simran K Arora, Paula Ribeiro, Sarah Kimmins, and Jianguo Xia. "miRNet-dissecting miRNA-target interactions and functional associations through network-based visual analysis." In: *Nucleic acids research* 44.W1 (2016), W135–W141.

[144] Yiwen Fang and Melissa J Fullwood. "Roles, functions, and mechanisms of long non-coding RNAs in cancer." In: *Genomics, proteomics & bioinformatics* 14.1 (2016), pp. 42–54.

[145] Antonella Fara, Zan Mitrev, Rodney Alexander Rosalia, and Bakri M Assas. "Cytokine storm and COVID-19: a chronicle of pro-inflammatory cytokines." In: *Open biology* 10.9 (2020), p. 200160.

[146] Dávid Fazekas, Mihály Koltai, Dénes Türei, Dezső Módos, Máté Pálfy, Zoltán Dúl, Lilian Zsákai, Máté Szalay-Bekő, Katalin Lenti, Illés J Farkas, et al. "SignaLink 2–a signaling pathway resource with multi-layered regulatory networks." In: *BMC systems biology* 7.1 (2013), pp. 1–15.

[147] Gary Felsenfeld and Mark Groudine. "Controlling the double helix." In: *Nature* 421.6921 (2003), pp. 448–453.

[148] Christopher J Ferrante and Samuel Joseph Leibovich. "Regulation of macrophage polarization and wound healing." In: *Advances in wound care* 1.1 (2012), pp. 10–16.

[149] Roy Fielding, Jim Gettys, Jeffrey Mogul, Henrik Frystyk, Larry Masinter, Paul Leach, and Tim Berners-Lee. *Hypertext transfer protocol–HTTP/1.1*. 1999.

[150] Roy T Fielding. *Architectural styles and the design of network-based software architectures*. Vol. 7. University of California, Irvine Irvine, 2000.

[151] Christine M Fillmore, Piyush B Gupta, Jenny A Rudnick, Silvia Caballero, Patricia J Keller, Eric S Lander, and Charlotte Kuperwasser. "Estrogen expands breast cancer stem-like cells through paracrine FGF/Tbx3 signaling." In: *Proceedings of the National Academy of Sciences* 107.50 (2010), pp. 21737–21742.

[152] Helmut Finner. "On a monotonicity problem in step-down multiple test procedures." In: *Journal of the American Statistical Association* 88.423 (1993), pp. 920–923.

[153] Rosie Fisher, Lazos Pusztai, and C Swanton. "Cancer heterogeneity: implications for targeted therapeutics." In: *British journal of cancer* 108.3 (2013), pp. 479–485.

[154] Simon Fishilevich, Ron Nudel, Noa Rappaport, Rotem Hadar, Inbar Plaschkes, Tsippi Iny Stein, Naomi Rosen, Asher Kohn, Michal Twik, Marilyn Safran, et al. "GeneHancer: genome-wide integration of enhancers and target genes in GeneCards." In: *Database* 2017 (2017).

[155] Alexandra Flemming. "Finding the perfect combination." In: *Nature Reviews Drug Discovery* 14.1 (2015), pp. 13–13.

[156] Michael NC Fletcher, Mauro AA Castro, Xin Wang, Ines De Santiago, Martin O'Reilly, Suet-Feung Chin, Oscar M Rueda, Carlos Caldas, Bruce AJ Ponder, Florian Markowetz, et al. "Master regulators of FGFR2 signalling and breast cancer risk." In: *Nature communications* 4.1 (2013), pp. 1–12.

[157] Simon A Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, et al. "COSMIC: somatic cancer genetics at high-resolution." In: *Nucleic acids research* 45.D1 (2017), pp. D777–D783.

[158] Oriol Fornes, Jaime A Castro-Mondragon, Aziz Khan, Robin Van der Lee, Xi Zhang, Phillip A Richmond, Bhavi P Modi, Solenne Correard, Marius Gheorghe, Damir Baranašić, et al. "JASPAR 2020: update of the open-access database of transcription factor binding profiles." In: *Nucleic acids research* 48.D1 (2020), pp. D87–D92.

[159] Eclipse Foundation. *Jakarta RESTful Web Services*. URL: https://jakarta.ee/specifications/restful-ws/ (visited on 03/11/2021).

[160] Eclipse Foundation. *Jakarta Servlet*. URL: https://jakarta.ee/specifications/servlet/ (visited on 03/11/2021).

[161] OpenJS Foundation. *jQuery*. URL: https://jquery.com/ (visited on 03/11/2021).

[162] Max Franz, Christian T Lopes, Gerardo Huck, Yue Dong, Onur Sumer, and Gary D Bader. "Cytoscape. js: a graph theory library for visualisation and analysis." In: *Bioinformatics* 32.2 (2016), pp. 309–311.

[163] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001.

[164] Claude Caron De Fromentel and Thierry Soussi. "TP53 tumor suppressor gene: a model for investigating human mutagenesis." In: *Genes, chromosomes and cancer* 4.1 (1992), pp. 1–15.

[165] R Fukuzawa, MR Anaka, RJ Weeks, IM Morison, and AE Reeve. "Canonical WNT signalling determines lineage specificity in Wilms tumour." In: *Oncogene* 28.8 (2009), pp. 1063–1075.

[166] R Furtwängler, N Nourkami, M Alkassar, D von Schweinitz, J-P Schenk, C Rübe, S Siemer, I Leuschner, and N Graf. "Update on relapses in unilateral nephroblastoma registered in 3 consecutive SIOP/GPOH studies–a report from the GPOH-nephroblastoma study group." In: *Klinische Pädiatrie* 223.03 (2011), pp. 113–119.

[167] Francis Galton. "Regression towards mediocrity in hereditary stature." In: *The Journal of the Anthropological Institute of Great Britain and Ireland* 15 (1886), pp. 246–263.

[168] Erich Gamma, Richard Helm, Ralph Johnson, John Vlissides, and Design Patterns. "Elements of Reusable Object-Oriented Software." In: *Design Patterns. massachusetts: Addison-Wesley Publishing Company* (1995).

[169] Monica Gandhi, Deborah S Yokoe, and Diane V Havlir. *Asymptomatic transmission, the Achilles' heel of current strategies to control Covid-19*. 2020.

[170] Tomas Ganz and Robert I Lehrer. "Defensins." In: *Current opinion in immunology* 6.4 (1994), pp. 584–589.

[171] Charles Garnier. *Les filaments basaux des cellules glandulaires. Note préliminaire*. Berger-Levrault, 1897.

[172] Jesse James Garrett et al. "Ajax: A new approach to web applications." In: (2005).

[173] Miklos Gaszner and Gary Felsenfeld. "Insulators: exploiting transcriptional and epigenetic mechanisms." In: *Nature Reviews Genetics* 7.9 (2006), pp. 703–713.

[174] Jens Geginat, Moira Paroni, Stefano Maglie, Johanna Sophie Alfen, Ilko Kastirr, Paola Gruarin, Marco De Simone, Massimiliano Pagani, and Sergio Abrignani. "Plasticity of human CD4 T cell subsets." In: *Frontiers in immunology* 5 (2014), p. 630.

[175] Andreas Gerasch, Daniel Faber, Jan Küntzer, Peter Niermann, Oliver Kohlbacher, Hans-Peter Lenhof, and Michael Kaufmann. "BiNA: a visual analytics tool for biological network data." In: *PloS one* 9.2 (2014), e87397.

[176] Nico Gerstner, **Tim Kehl**, Kerstin Lenhof, Lea Eckhart, Lara Schneider, Daniel Stöckel, Christina Backes, Eckart Meese, Andreas Keller, and Hans-Peter Lenhof. "GeneTrail: a framework for the analysis of high-throughput profiles." In: *Frontiers in Molecular Biosciences* (2021), p. 890.

[177] Nico Gerstner, **Tim Kehl**, Kerstin Lenhof, Anne Müller, Carolin Mayer, Lea Eckhart, Nadja Liddy Grammes, Caroline Diener, Martin Hart, Oliver Hahn, et al. "GeneTrail 3: advanced high-throughput enrichment analysis." In: *Nucleic Acids Research* (2020).

[178] Evangelos J Giamarellos-Bourboulis, Mihai G Netea, Nikoletta Rovina, Karolina Akinosoglou, Anastasia Antoniadou, Nikolaos Antonakos, Georgia Damoraki, Theologia Gkavogianni, Maria-Evangelia Adami, Paraskevi Katsaounou, et al. "Complex immune dysregulation in COVID-19 patients with severe respiratory failure." In: *Cell host & microbe* 27.6 (2020), pp. 992–1000.

[179] Gary L Glish and Richard W Vachet. "The basics of mass spectrometry in the twenty-first century." In: *Nature reviews drug discovery* 2.2 (2003), pp. 140–150.

[180] CeGaT GmbH. *CeGaT*. 2021. URL: https://www.cegat.de (visited on 11/25/2021).

[181] Jeremy Goecks, Anton Nekrutenko, and James Taylor. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." In: *Genome Biol* 11.8 (2010), R86.

[182] Matthew P Goetz. "Tamoxifen, endoxifen, and CYP2D6: the rules for evaluating a predictive factor." In: *Oncology* 23.14 (2009).

[183] Joana P Gonçalves, Ricardo S Aires, Alexandre P Francisco, and Sara C Madeira. "Regulatory snapshots: integrative mining of regulatory modules from expression time series and regulatory networks." In: *PloS one* 7.5 (2012), e35977.

[184] Joana P Gonçalves, Alexandre P Francisco, Nuno P Mira, Miguel C Teixeira, Isabel Sá-Correia, Arlindo L Oliveira, and Sara C Madeira. "TFRank: network-based prioritization of regulatory associations underlying transcriptional responses." In: *Bioinformatics* 27.22 (2011), pp. 3149–3157.

[185] Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, David Tamborero, Michael P Schroeder, Alba Jene-Sanz, Alberto Santos, and Nuria Lopez-Bigas. "IntOGen-mutations identifies cancer drivers across tumor types." In: *Nature methods* 10.11 (2013), pp. 1081–1082.

[186] Aaron M Goodman, Shumei Kato, Lyudmila Bazhenova, Sandip P Patel, Garrett M Frampton, Vincent Miller, Philip J Stephens, Gregory A Daniels, and Razelle Kurzrock. "Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers." In: *Molecular cancer therapeutics* 16.11 (2017), pp. 2598–2608.

[187] Siamon Gordon. *The macrophage as therapeutic target*. Vol. 158. Springer Science & Business Media, 2012.

[188] Norbert Graf, Marie-France Tournade, and Jan de Kraker. "The role of preoperative chemotherapy in the management of Wilms' tumor: The SIOP Studies." In: *Urologic Clinics of North America* 27.3 (2000), pp. 443–454.

[189] Carla Grandori, Shaun M Cowley, Leonard P James, and Robert N Eisenman. "The Myc/Max/Mad network and the transcriptional control of cell behavior." In: *Annual review of cell and developmental biology* 16.1 (2000), pp. 653–699.

[190] Charles E Grant, Timothy L Bailey, and William Stafford Noble. "FIMO: scanning for occurrences of a given motif." In: *Bioinformatics* 27.7 (2011), pp. 1017–1018.

[191] Malachi Griffith, Nicholas C Spies, Kilannin Krysiak, Joshua F McMichael, Adam C Coffman, Arpad M Danos, Benjamin J Ainscough, Cody A Ramirez, Damian T Rieke, Lynzey Kujan, et al. "CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer." In: *Nature genetics* 49.2 (2017), pp. 170–174.

[192] Martin Grötschel and Yoshiko Wakabayashi. "A cutting plane algorithm for a clustering problem." In: *Mathematical Programming* 45.1 (1989), pp. 59–96.

[193] Adrienne Grzenda, Gwen Lomberk, Jin-San Zhang, and Raul Urrutia. "Sin3: master scaffold and transcriptional corepressor." In: *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1789.6-8 (2009), pp. 443–450.

[194] DP Guimaraes and P Hainaut. "TP53: a key gene in human cancer." In: *Biochimie* 84.1 (2002), pp. 83–93.

[195] Chuang Guo, Bin Li, Huan Ma, Xiaofang Wang, Pengfei Cai, Qiaoni Yu, Lin Zhu, Liying Jin, Chen Jiang, Jingwen Fang, et al. "Single-cell analysis of two severe COVID-19 patients reveals a monocyte-associated and tocilizumab-responding cytokine storm." In: *Nature communications* 11.1 (2020), pp. 1–11.

[196] Eduardo G Gusmao, Manuel Allhoff, Martin Zenke, and Ivan G Costa. "Analysis of computational footprinting methods for DNase sequencing experiments." In: *Nature methods* 13.4 (2016), pp. 303–309.

[197] Eduardo G Gusmao, Christoph Dieterich, Martin Zenke, and Ivan G Costa. "Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications." In: *Bioinformatics* 30.22 (2014), pp. 3143–3151.

[198] Jérôme Hadjadj, Nader Yatim, Laura Barnabei, Aurélien Corneau, Jeremy Boussier, Nikaïa Smith, Hélène Péré, Bruno Charbit, Vincent Bondet, Camille Chenevier-Gobeaux, et al. "Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients." In: *Science* 369.6504 (2020), pp. 718–724.

[199] Sevil Oskay Halacli and Ayse Lale Dogan. "FOXP1 regulation via the PI3K/Akt/p70S6K signaling pathway in breast cancer cells." In: *Oncology letters* 9.3 (2015), pp. 1482–1488.

[200] Richard W Hamming. "Error detecting and error correcting codes." In: *The Bell system technical journal* 29.2 (1950), pp. 147–160.

[201] Douglas Hanahan and Robert A Weinberg. "The hallmarks of cancer." In: *cell* 100.1 (2000), pp. 57–70.

[202] Douglas Hanahan and Robert A Weinberg. "Hallmarks of cancer: the next generation." In: *cell* 144.5 (2011), pp. 646–674.

[203] John M Hancock. "HAVANA (Human and Vertebrate Analysis and Annotation)." In: *Dictionary of Bioinformatics and Computational Biology* (2004).

[204] Yuhan Hao et al. "Integrated analysis of multimodal single-cell data." In: *Cell* (2021). DOI: 10.1016/j.cell.2021.04.048. URL: https://doi.org/10.1016/j.cell.2021.04.048.

[205] Alan W Harris, CARL A Pinkert, M Crawford, WY Langdon, RL Brinster, and JM Adams. "The E mu-myc transgenic mouse. A model for high-incidence spontaneous lymphoma and leukemia of early B cells." In: *Journal of Experimental Medicine* 167.2 (1988), pp. 353–371.

[206] C Jake Harris, Marion Scheibe, Somsakul Pop Wongpalee, Wanlu Liu, Evan M Cornett, Robert M Vaughan, Xueqin Li, Wei Chen, Yan Xue, Zhenhui Zhong, et al. "A DNA methylation reader complex that enhances gene transcription." In: *Science* 362.6419 (2018), pp. 1182–1186.

[207] Jennifer Harrow, France Denoeud, Adam Frankish, Alexandre Reymond, Chao-Kung Chen, Jacqueline Chrast, Julien Lagarde, James GR Gilbert, Roy Storey, David Swarbreck, et al. "GENCODE: producing a reference annotation for ENCODE." In: *Genome biology* 7.1 (2006), pp. 1–9.

[208] Tibshirani Hastie and Robert Tibshirani. *& Friedman, J.(2008). The Elements of Statistical Learning; Data Mining, Inference and Prediction*. 2009.

[209] Housheng Hansen He, Clifford A Meyer, Mei-Wei Chen, Chongzhi Zang, Yin Liu, Prakash K Rao, Teng Fei, Han Xu, Henry Long, X Shirley Liu, et al. "Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification." In: *Nature methods* 11.1 (2014), p. 73.

[210] Laura M Heiser, Anguraj Sadanandam, Wen-Lin Kuo, Stephen C Benz, Theodore C Goldstein, Sam Ng, William J Gibb, Nicholas J Wang, Safiyyah Ziyad, Frances Tong, et al. "Subtype and pathway specific responses to anticancer compounds in breast cancer." In: *Proceedings of the National Academy of Sciences* 109.8 (2012), pp. 2724–2729.

[211] Micheal Hewett, Diane E Oliver, Daniel L Rubin, Katrina L Easton, Joshua M Stuart, Russ B Altman, and Teri E Klein. "PharmGKB: the pharmacogenetics knowledge base." In: *Nucleic acids research* 30.1 (2002), pp. 163–165.

[212] Denes Hnisz, Brian J Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-André, Alla A Sigova, Heather A Hoke, and Richard A Young. "Super-enhancers in the control of cell identity and disease." In: *Cell* 155.4 (2013), pp. 934–947.

[213] Suchita Hodawadekar, Duonan Yu, Diana Cozma, Bruce Freedman, Oriol Sunyer, Michael L Atchison, and Andrei Thomas-Tikhonenko. "B-lymphoma cells with epigenetic silencing of Pax5 trans-differentiate into macrophages, but not other hematopoietic lineages." In: *Experimental cell research* 313.2 (2007), pp. 331–340.

[214] Jan HJ Hoeijmakers. "Genome maintenance mechanisms for preventing cancer." In: *nature* 411.6835 (2001), pp. 366–374.

[215] Robin Holliday. "DNA methylation and epigenetic mechanisms." In: *Cell biophysics* 15.1-2 (1989), pp. 15–20.

[216] Katherine Hollywood, Daniel R Brison, and Royston Goodacre. "Metabolomics: current technologies and future trends." In: *Proteomics* 6.17 (2006), pp. 4716–4723.

[217] Sture Holm. "A simple sequentially rejective multiple test procedure." In: *Scandinavian journal of statistics* (1979), pp. 65–70.

[218] Nils Homer, Barry Merriman, and Stanley F Nelson. "BFAST: an alignment tool for large scale genome resequencing." In: *PloS one* 4.11 (2009), e7767.

[219] Robert Hooke. *Micrographia (1665)*. Vol. 20. AppLife, 2014.

[220] Yu Hou, Huahu Guo, Chen Cao, Xianlong Li, Boqiang Hu, Ping Zhu, Xinglong Wu, Lu Wen, Fuchou Tang, Yanyi Huang, et al. "Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas." In: *Cell research* 26.3 (2016), pp. 304–319.

[221] David C Hoyle, Magnus Rattray, Ray Jupp, and Andrew Brass. "Making sense of microarray data distributions." In: *Bioinformatics* 18.4 (2002), pp. 576–584.

[222] Justin Bo-Kai Hsu, Chih-Min Chiu, Sheng-Da Hsu, Wei-Yun Huang, Chia-Hung Chien, Tzong-Yi Lee, and Hsien-Da Huang. "miRTar: an integrated system for identifying miRNA-target interactions in human." In: *BMC bioinformatics* 12.1 (2011), pp. 1–12.

[223] Gongzhu Hu, Jinping Wang, and Wenying Feng. "Multivariate regression modeling for home value estimates with evaluation using maximum information coefficient." In: *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing 2012*. Springer, 2013, pp. 69–81.

[224] Ping Hu, Wenhua Zhang, Hongbo Xin, and Glenn Deng. "Single cell isolation and analysis." In: *Frontiers in cell and developmental biology* 4 (2016), p. 116.

[225] Chia-Ling Huang, John Lamb, Leonid Chindelevitch, Jarek Kostrowicki, Justin Guinney, Charles DeLisi, and Daniel Ziemek. "Correlation set analysis: detecting active regulators in disease populations using prior causal knowledge." In: *BMC bioinformatics* 13.1 (2012), pp. 1–15.

[226] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." In: *Nucleic acids research* 37.1 (2009), pp. 1–13.

[227] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." In: *Nature protocols* 4.1 (2009), pp. 44–57.

[228] Hsi-Yuan Huang, Yang-Chi-Dung Lin, Jing Li, Kai-Yao Huang, Sirjana Shrestha, Hsiao-Chin Hong, Yun Tang, Yi-Gang Chen, Chen-Nan Jin, Yuan Yu, et al. "miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database." In: *Nucleic acids research* 48.D1 (2020), pp. D148–D154.

[229] Wolfgang Huber, Anja Von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron. "Variance stabilization applied to microarray data calibration and to the quantification of differential expression." In: *Bioinformatics* 18.suppl_1 (2002), S96–S104.

[230] Nicholas J Hudson, Antonio Reverter, and Brian P Dalrymple. "A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation." In: *PLoS Comput Biol* 5.5 (2009), e1000382.

[231] Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. "Gene set enrichment analysis: performance evaluation and usage guidelines." In: *Briefings in bioinformatics* 13.3 (2012), pp. 281–291.

[232] Motoshi Ichikawa, Akihide Yoshimi, Masahiro Nakagawa, Nahoko Nishimoto, Naoko Watanabe-Okochi, and Mineo Kurokawa. "A role for RUNX1 in hematopoiesis and myeloid leukemia." In: *International journal of hematology* 97.6 (2013), pp. 726–734.

[233] Illumina. *An introduction to Next-Generation Sequencing Technology*. 2020. URL: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf (visited on 10/09/2020).

[234] Illumina. *HiSeq™ Sequencing Systems*. 2020. URL: https://www.illumina.com/documents/products/datasheets/datasheet_hiseq_systems.pdf (visited on 12/26/2020).

[235] Illumina. *llumina sequencing platforms*. 2020. URL: https://www.illumina.com/systems/sequencing-platforms.html (visited on 10/20/2020).

[236] Illumina. *Infinium Methylation Assay*. 2022. URL: https://www.illumina.com/science/technology/microarray/infinium-methylation-assay.html (visited on 02/01/2022).

[237] Broad Institute. *GSEA - Data formats*. 2022. URL: https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats (visited on 04/21/2022).

[238] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, et al. "A landscape of pharmacogenomic interactions in cancer." In: *Cell* 166.3 (2016), pp. 740–754.

[239] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." In: *Biostatistics* 4.2 (2003), pp. 249–264.

[240] Paul Jaccard. "The distribution of the flora in the alpine zone. 1." In: *New phytologist* 11.2 (1912), pp. 37–50.

[241] Benoît Jacob, Gaël Guennebaud, et al. *Eigen 3*. URL: https://eigen.tuxfamily.org/ (visited on 04/17/2021).

[242] Charles A Janeway Jr, Paul Travers, Mark Walport, and Mark J Shlomchik. "The complement system and innate immunity." In: *Immunobiology: The Immune System in Health and Disease. 5th edition*. Garland Science, 2001.

[243] Nico AF Janssen, Inge Grondman, Aline H de Nooijer, Collins K Boahen, Valerie ACM Koeken, Vasiliki Matzaraki, Vinod Kumar, Xuehui He, Matthijs Kox, Hans JPM Koenen, et al. "Dysregulated innate and adaptive immune responses discriminate disease severity in COVID-19." In: *The Journal of Infectious Diseases* 223.8 (2021), pp. 1322–1333.

[244] Thomas Jenuwein and C David Allis. "Translating the histone code." In: *Science* 293.5532 (2001), pp. 1074–1080.

[245] Wenfei Jin, Qingsong Tang, Mimi Wan, Kairong Cui, Yi Zhang, Gang Ren, Bing Ni, Jeffrey Sklar, Teresa M Przytycka, Richard Childs, et al. "Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples." In: *Nature* 528.7580 (2015), pp. 142–146.

[246] Stefanie Jonas and Elisa Izaurralde. "Towards a molecular understanding of microRNA-mediated gene silencing." In: *Nature reviews genetics* 16.7 (2015), pp. 421–433.

[247] Peter A Jones and Daiya Takai. "The role of DNA methylation in mammalian epigenetics." In: *Science* 293.5532 (2001), pp. 1068–1070.

[248] David Juan, Juliane Perner, Enrique Carrillo de Santa Pau, Simone Marsili, David Ochoa, Ho-Ryun Chung, Martin Vingron, Daniel Rico, and Alfonso Valencia. "Epigenomic co-localization and co-evolution reveal a key role for 5hmC as a communication hub in the chromatin network of ESCs." In: *Cell reports* 14.5 (2016), pp. 1246–1257.

[249] Juhani Kähärä and Harri Lähdesmäki. "BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data." In: *Bioinformatics* 31.17 (2015), pp. 2852–2859.

[250] Raghu Kalluri, Robert A Weinberg, et al. "The basics of epithelial-mesenchymal transition." In: *The Journal of clinical investigation* 119.6 (2009), pp. 1420–1428.

[251] Atanas Kamburov, Christoph Wierling, Hans Lehrach, and Ralf Herwig. "ConsensusPathDB—a database for integrating human functional interaction networks." In: *Nucleic acids research* 37.suppl_1 (2009), pp. D623–D628.

[252] Gopal K Kanji. *100 statistical tests*. Sage, 2006.

[253] Dimitra Karagkouni, Maria D Paraskevopoulou, Serafeim Chatzopoulos, Ioannis S Vlachos, Spyros Tastsoglou, Ilias Kanellos, Dimitris Papadimitriou, Ioannis Kavakiotis, Sofia Maniou, Giorgos Skoufos, et al. "DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions." In: *Nucleic acids research* 46.D1 (2018), pp. D239–D245.

[254] Kiyokazu Kawabe, Daniel Lindsay, Manjit Braitch, Angela J Fahey, Louise Showe, and Cris S Constantinescu. "IL-12 inhibits glucocorticoid-induced T cell apoptosis by inducing GMEB1 and activating PI3K/Akt pathway." In: *Immunobiology* 217.1 (2012), pp. 118–123.

[255] Bridget Keenen and Ivana L de la Serna. "Chromatin remodeling in embryonic stem cells: regulating the balance between pluripotency and differentiation." In: *Journal of cellular physiology* 219.1 (2009), pp. 1–7.

[256] Thomas Kelder, Martijn P Van Iersel, Kristina Hanspers, Martina Kutmon, Bruce R Conklin, Chris T Evelo, and Alexander R Pico. "WikiPathways: building research communities on biological pathways." In: *Nucleic acids research* 40.D1 (2012), pp. D1301–D1307.

[257] Andreas Keller, Christina Backes, Maher Al-Awadhi, Andreas Gerasch, Jan Küntzer, Oliver Kohlbacher, Michael Kaufmann, and Hans-Peter Lenhof. "GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments." In: *BMC bioinformatics* 9.1 (2008), pp. 1–6.

[258] Andreas Keller, Christina Backes, Andreas Gerasch, Michael Kaufmann, Oliver Kohlbacher, Eckart Meese, and Hans-Peter Lenhof. "A novel algorithm for detecting differentially regulated paths based on gene set enrichment analysis." In: *Bioinformatics* 25.21 (2009), pp. 2787–2794.

[259] Andreas Keller, Christina Backes, and Hans-Peter Lenhof. "Computation of significance scores of unweighted Gene Set Enrichment Analyses." In: *BMC bioinformatics* 8.1 (2007), pp. 1–7.

[260] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. "The human genome browser at UCSC." In: *Genome research* 12.6 (2002), pp. 996–1006.

[261] Fabian Kern, Ernesto Aparicio-Puerta, Yongping Li, Tobias Fehlmann, **Tim Kehl**, Viktoria Wagner, Kamalika Ray, Nicole Ludwig, Hans-Peter Lenhof, Eckart Meese, et al. "miRTargetLink 2.0—interactive miRNA target gene and target pathway networks." In: *Nucleic Acids Research* (2021).

[262] Satyajeet P Khare, Farhat Habib, Rahul Sharma, Nikhil Gadewal, Sanjay Gupta, and Sanjeev Galande. "HIstome—a relational knowledgebase of human histone proteins and histone modifying enzymes." In: *Nucleic acids research* 40.D1 (2012), pp. D337–D342.

[263] Budiman Kharma, Tsukasa Baba, Noriomi Matsumura, Hyun Sook Kang, Junzo Hamanishi, Ryusuke Murakami, Melissa M McConechy, Samuel Leung, Ken Yamaguchi, Yuko Hosoe, et al. "STAT1 drives tumor progression in serous papillary endometrial cancer." In: *Cancer research* 74.22 (2014), pp. 6519–6530.

[264] Purvesh Khatri, Marina Sirota, and Atul J Butte. "Ten years of pathway analysis: current approaches and outstanding challenges." In: *PLoS Comput Biol* 8.2 (2012), e1002375.

[265] Pouya Kheradpour and Manolis Kellis. "Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments." In: *Nucleic acids research* 42.5 (2014), pp. 2976–2987.

[266] Kyong-Rim Kieffer-Kwon, Keisuke Nimura, Suhas SP Rao, Jianliang Xu, Seolkyoung Jung, Aleksandra Pekowska, Marei Dose, Evan Stevens, Ewy Mathe, Peng Dong, et al. "Myc regulates chromatin decompaction and nuclear architecture during B cell activation." In: *Molecular cell* 67.4 (2017), pp. 566–578.

[267] Nam W Kim, Mieczyslaw A Piatyszek, Karen R Prowse, Calvin B Harley, Michael D West, P d L Ho, Gina M Coviello, Woodring E Wright, Scott L Weinrich, and Jerry W Shay. "Specific association of human telomerase activity with immortal cells and cancer." In: *Science* 266.5193 (1994), pp. 2011–2015.

[268] Hiroshi Kimura. "Histone modifications for human epigenome analysis." In: *Journal of human genetics* 58.7 (2013), pp. 439–445.

[269] Rhoda J Kinsella, Andreas Kähäri, Syed Haider, Jorge Zamora, Glenn Proctor, Giulietta Spudich, Jeff Almeida-King, Daniel Staines, Paul Derwent, Arnaud Kerhornou, et al. "Ensembl BioMarts: a hub for data retrieval across taxonomic space." In: *Database* 2011 (2011).

[270] Sandy L Klemm, Zohar Shipony, and William J Greenleaf. "Chromatin accessibility and the regulatory epigenome." In: *Nature Reviews Genetics* 20.4 (2019), pp. 207–220.

[271] Theo A Knijnenburg, Lodewyk FA Wessels, Marcel JT Reinders, and Ilya Shmulevich. "Fewer permutations, more accurate P-values." In: *Bioinformatics* 25.12 (2009), pp. i161–i168.

[272] Paul S Knoepfler, Xiao-yong Zhang, Pei Feng Cheng, Philip R Gafken, Steven B McMahon, and Robert N Eisenman. "Myc influences global chromatin structure." In: *The EMBO journal* 25.12 (2006), pp. 2723–2734.

[273] DCFR Koboldt, Robert Fulton, Michael McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua McMichael, Lucinda Fulton, David Dooling, Li Ding, Elaine Mardis, et al. "Comprehensive molecular portraits of human breast tumours." In: *Nature* 490.7418 (2012), pp. 61–70.

[274] Andrey Kolmogorov. "Sulla determinazione empirica di una lgge di distribuzione." In: *Inst. Ital. Attuari, Giorn.* 4 (1933), pp. 83–91.

[275] Say Li Kong, Guoliang Li, Siang Lin Loh, Wing-Kin Sung, and Edison T Liu. "Cellular reprogramming by the conjoint action of ERα, FOXA1, and GATA3 to a ligand-inducible growth state." In: *Molecular systems biology* 7.1 (2011), p. 526.

[276] Sergey Koren and Adam M Phillippy. "One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly." In: *Current opinion in microbiology* 23 (2015), pp. 110–120.

[277] Esther Korpershoek, Djamailys Koffy, Bert H Eussen, Lindsey Oudijk, Thomas G Papathomas, Francien H van Nederveen, Eric JT Belt, Gaston JH Franssen, David FJ Restuccia, Niels MG Krol, et al. "Complex MAX rearrangement in a family with malignant pheochromocytoma, renal oncocytoma, and erythrocytosis." In: *The Journal of Clinical Endocrinology & Metabolism* 101.2 (2016), pp. 453–460.

[278] Keegan Korthauer, Patrick K Kimes, Claire Duvallet, Alejandro Reyes, Ayshwarya Subramanian, Mingxiang Teng, Chinmay Shukla, Eric J Alm, and Stephanie C Hicks. "A practical guide to methods controlling false discoveries in computational biology." In: *Genome biology* 20.1 (2019), pp. 1–21.

[279] Michael R Kosorok. "Discussion of: Brownian distance covariance." In: *The Annals of Applied Statistics* 3.4 (2009), pp. 1270–1278.

[280] Jan Kotek et al. *MapDB*. URL: https://mapdb.org/ (visited on 03/04/2021).

[281] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. "miRBase: from microRNA sequences to function." In: *Nucleic acids research* 47.D1 (2019), pp. D155–D162.

[282] Mel N Kronick. "Creation of the whole human genome microarray." In: *Expert review of proteomics* 1.1 (2004), pp. 19–28.

[283] Ivan V Kulakovskiy, Ilya E Vorontsov, Ivan S Yevshin, Ruslan N Sharipov, Alla D Fedorova, Eugene I Rumynskiy, Yulia A Medvedeva, Arturo Magana-Mora, Vladimir B Bajic, Dmitry A Papatsenko, et al. "HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis." In: *Nucleic acids research* 46.D1 (2018), pp. D252–D259.

[284] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update." In: *Nucleic acids research* 44.W1 (2016), W90–W97.

[285] Prateek Kumar, Steven Henikoff, and Pauline C Ng. "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm." In: *Nature protocols* 4.7 (2009), pp. 1073–1081.

[286] Alexander Lachmann, Huilei Xu, Jayanth Krishnan, Seth I Berger, Amin R Mazloom, and Avi Ma'ayan. "ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments." In: *Bioinformatics* 26.19 (2010), pp. 2438–2444.

[287] Hichem Lahouassa, Marie-Lise Blondot, Lise Chauveau, Ghina Chougui, Marina Morel, Marjorie Leduc, François Guillonneau, Bertha Cecilia Ramirez, Olivier Schwartz, and Florence Margottin-Goguet. "HIV-1 Vpr degrades the HLTF DNA translocase in T cells and macrophages." In: *Proceedings of the National Academy of Sciences* 113.19 (2016), pp. 5311–5316.

[288] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. "Initial sequencing and analysis of the human genome." In: (2001).

[289] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2." In: *Nature methods* 9.4 (2012), p. 357.

[290] M Lauss et al. *Mutational and putative neoantigen load predict clinical benefit of adoptive T cell therapy in melanoma. Nat Commun 8, 1738*. 2017.

[291] Devon A Lawson, Nirav R Bhakta, Kai Kessenbrock, Karin D Prummel, Ying Yu, Ken Takai, Alicia Zhou, Henok Eyob, Sanjeev Balakrishnan, Chih-Yang Wang, et al. "Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells." In: *Nature* 526.7571 (2015), pp. 131–135.

[292] Jeong Seok Lee and Eui-Cheol Shin. "The type I interferon response in COVID-19: implications for treatment." In: *Nature Reviews Immunology* 20.10 (2020), pp. 585–586.

[293] Tong Ihn Lee and Richard A Young. "Transcriptional regulation and its misregulation in disease." In: *Cell* 152.6 (2013), pp. 1237–1251.

[294] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. "Tackling the widespread and critical impact of batch effects in high-throughput data." In: *Nature Reviews Genetics* 11.10 (2010), pp. 733–739.

[295] Gustavo Leone, Rosalie Sears, Erich Huang, Rachel Rempel, Faison Nuckolls, Chi-Hyun Park, Paloma Giangrande, Lizhao Wu, Harold I Saavedra, Seth J Field, et al. "Myc requires distinct E2F activities to induce S phase and apoptosis." In: *Molecular cell* 8.1 (2001), pp. 105–113.

[296] Chi-Ming Li, Meirong Guo, Alain Borczuk, Charles A Powell, Michelle Wei, Harshwardhan M Thaker, Richard Friedman, Ulf Klein, and Benjamin Tycko. "Gene expression in Wilms' tumor mimics the earliest committed stage in the metanephric mesenchymal-epithelial transition." In: *The American journal of pathology* 160.6 (2002), pp. 2181–2190.

[297] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph." In: *Bioinformatics* 31.10 (2015), pp. 1674–1676.

[298] Guoqiang Li, Yaping Liu, Yanxiao Zhang, Naoki Kubo, Miao Yu, Rongxin Fang, Manolis Kellis, and Bing Ren. "Joint profiling of DNA methylation and chromatin architecture in single cells." In: *Nature methods* 16.10 (2019), pp. 991–993.

[299] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform." In: *bioinformatics* 25.14 (2009), pp. 1754–1760.

[300] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. "The sequence alignment/map format and SAMtools." In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[301] Jun Li, Daniela M Witten, Iain M Johnstone, and Robert Tibshirani. "Normalization, testing, and false discovery rate estimation for RNA-sequencing data." In: *Biostatistics* 13.3 (2012), pp. 523–538.

[302] Ruiqiang Li, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, Shengting Li, Gao Shan, Karsten Kristiansen, et al. "De novo assembly of human genomes with massively parallel short read sequencing." In: *Genome research* 20.2 (2010), pp. 265–272.

[303] Chunguang Liang, Elena Bencurova, Eric Psota, Priya Neurgaonkar, Martina Prelog, Carsten Scheller, and Thomas Dandekar. "Population-Predicted MHC Class II Epitope Presentation of SARS-CoV-2 Structural Proteins Correlates to the Case Fatality Rates of COVID-19 in Different Countries." In: *International journal of molecular sciences* 22.5 (2021), p. 2630.

[304] Xudong Liao, Nikunj Sharma, Fehmida Kapadia, Guangjin Zhou, Yuan Lu, Hong Hong, Kaavya Paruchuri, Ganapati H Mahabeleshwar, Elise Dalmas, Nicolas Venteclef, et al. "Krüppel-like factor 4 regulates macrophage polarization." In: *The Journal of clinical investigation* 121.7 (2011), pp. 2736–2749.

[305] Yuxing Liao, Jing Wang, Eric J Jaehnig, Zhiao Shi, and Bing Zhang. "WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs." In: *Nucleic acids research* 47.W1 (2019), W199–W205.

[306] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. "Molecular signatures database (MSigDB) 3.0." In: *Bioinformatics* 27.12 (2011), pp. 1739–1740.

[307] Astrid Lievre, Jean-Baptiste Bachet, Delphine Le Corre, Valerie Boige, Bruno Landi, Jean-François Emile, Jean-François Côté, Gorana Tomasic, Christophe Penna, Michel Ducreux, et al. "KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer." In: *Cancer research* 66.8 (2006), pp. 3992–3995.

[308] Gaye Lightbody, Valeriia Haberland, Fiona Browne, Laura Taggart, Huiru Zheng, Eileen Parkes, and Jaine K Blayney. "Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application." In: *Briefings in bioinformatics* 20.5 (2019), pp. 1795–1811.

[309] Geou-Yarh Liou and Peter Storz. "Reactive oxygen species in cancer." In: *Free radical research* 44.5 (2010), pp. 479–496.

[310] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. "Human DNA methylomes at base resolution show widespread epigenomic differences." In: *nature* 462.7271 (2009), pp. 315–322.

[311] Xiaojing Liu and Jason W Locasale. "Metabolomics: a primer." In: *Trends in biochemical sciences* 42.4 (2017), pp. 274–284.

[312] Google LLC. *Googletest - Google Testing and Mocking Framework*. URL: https://github.com/google/googletest (visited on 04/17/2021).

[313] Yuin-Han Loh, Qiang Wu, Joon-Lin Chew, Vinsensius B Vega, Weiwei Zhang, Xi Chen, Guillaume Bourque, Joshy George, Bernard Leong, Jun Liu, et al. "The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells." In: *Nature genetics* 38.4 (2006), pp. 431–440.

[314] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. "The genotype-tissue expression (GTEx) project." In: *Nature genetics* 45.6 (2013), pp. 580–585.

[315] Scott W Lowe, Enrique Cepero, and Gerard Evan. "Intrinsic tumour suppression." In: *Nature* 432.7015 (2004), pp. 307–315.

[316] SpryMedia Ltd. *DataTables*. URL: https://datatables.net/ (visited on 03/11/2021).

[317] Tzu-Pin Lu, Chien-Yueh Lee, Mong-Hsun Tsai, Yu-Chiao Chiu, Chuhsing Kate Hsiao, Liang-Chuan Lai, and Eric Y Chuang. "miRSystem: an integrated system for characterizing enriched functions and pathways of microRNA targets." In: (2012).

[318] Nicole Ludwig, Petra Leidinger, Kurt Becker, Christina Backes, Tobias Fehlmann, Christian Pallasch, Steffi Rheinheimer, Benjamin Meder, Cord Stähler, Eckart Meese, et al. "Distribution of miRNA expression across human tissues." In: *Nucleic acids research* 44.8 (2016), pp. 3865–3877.

[319] Nicole Ludwig, Tamara V Werner, Christina Backes, Patrick Trampert, Manfred Gessler, Andreas Keller, Hans-Peter Lenhof, Norbert Graf, and Eckart Meese. "Combining miRNA and mRNA expression profiles in Wilms tumor subtypes." In: *International journal of molecular sciences* 17.4 (2016), p. 475.

[320] Malte D Luecken and Fabian J Theis. "Current best practices in single-cell RNA-seq analysis: a tutorial." In: *Molecular systems biology* 15.6 (2019), e8746.

[321] Kaixuan Luo and Alexander J Hartemink. "Using DNase digestion data to accurately identify transcription factor binding sites." In: *Biocomputing 2013*. World Scientific, 2013, pp. 80–91.

[322] Feiyang Ma, Brie K Fuqua, Yehudit Hasin, Clara Yukhtman, Chris D Vulpe, Aldons J Lusis, and Matteo Pellegrini. "A comparison between whole transcript and 3'RNA sequencing methods using Kapa and Lexogen library preparation methods." In: *BMC genomics* 20.1 (2019), pp. 1–12.

[323] Xianwei Ma, Zhengyu Jiang, Na Li, Wei Jiang, Peng Gao, Mingjin Yang, Xiya Yu, Guifang Wang, and Yan Zhang. "Ets2 suppresses inflammatory cytokines through MAPK/NF-κB signaling and directly binds to the IL-6 promoter in macrophages." In: *Aging (Albany NY)* 11.22 (2019), p. 10610.

[324] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[325] Maria J Macias, Pau Martin-Malpartida, and Joan Massagué. "Structural determinants of Smad function in TGF-β signaling." In: *Trends in biochemical sciences* 40.6 (2015), pp. 296–308.

[326] Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. "Entrez Gene: gene-centered information at NCBI." In: *Nucleic acids research* 33.suppl_1 (2005), pp. D54–D58.

[327] Zuzanna Makowska, Tujana Boldanova, David Adametz, Luca Quagliata, Julia E Vogt, Michael T Dill, Mathias S Matter, Volker Roth, Luigi Terracciano, and Markus H Heim. "Gene expression analysis of biopsy samples reveals critical limitations of transcriptome-based molecular classifications of hepatocellular carcinoma." In: *The Journal of Pathology: Clinical Research* 2.2 (2016), pp. 80–92.

[328] Loïka Maltais, Martin Montagne, Mikaël Bédard, Cynthia Tremblay, Laura Soucek, and Pierre Lavigne. "Biophysical characterization of the b-HLH-LZ of ΔMax, an alternatively spliced isoform of Max found in tumor cells: Towards the validation of a tumor suppressor role for the Max homodimers." In: *PloS one* 12.3 (2017), e0174413.

[329] Thomas Manke, Helge G Roider, and Martin Vingron. "Statistical modeling of transcription factor binding affinities predicts regulatory interactions." In: *PLoS Comput Biol* 4.3 (2008), e1000039.

[330] CPLEX User's Manual. "Ibm ilog cplex optimization studio." In: *Version* 12 (1987), pp. 1987–2018.

[331] Santiago Marco-Sola, Michael Sammeth, Roderic Guigó, and Paolo Ribeca. "The GEM mapper: fast, accurate and versatile alignment by filtration." In: *Nature methods* 9.12 (2012), pp. 1185–1188.

[332] Kenneth Marek, Sohini Chowdhury, Andrew Siderowf, Shirley Lasch, Christopher S Coffey, Chelsea Caspell-Garcia, Tanya Simuni, Danna Jennings, Caroline M Tanner, John Q Trojanowski, et al. "The Parkinson's progression markers initiative (PPMI)–establishing a PD biomarker cohort." In: *Annals of clinical and translational neurology* 5.12 (2018), pp. 1460–1477.

[333] Alexander Marson, Stuart S Levine, Megan F Cole, Garrett M Frampton, Tobias Brambrink, Sarah Johnstone, Matthew G Guenther, Wendy K Johnston, Marius Wernig, Jamie Newman, et al. "Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells." In: *Cell* 134.3 (2008), pp. 521–533.

[334] Joost HA Martens and Hendrik G Stunnenberg. "BLUEPRINT: mapping human blood cell epigenomes." In: *Haematologica* 98.10 (2013), p. 1487.

[335] Ken Martin and Bill Hoffman. *CMake Reference Documentation*. URL: https://cmake.org/cmake/help/latest/index.html (visited on 04/17/2021).

[336] Ken Martin and Bill Hoffman. "An open source approach to developing software in a small organization." In: *IEEE software* 24.1 (2007), pp. 46–53.

[337] Andriy Marusyk and Kornelia Polyak. "Tumor heterogeneity: causes and consequences." In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1805.1 (2010), pp. 105–117.

[338] Mariana Maschietto, Fabio S Piccoli, Cecilia ML Costa, Luiz P Camargo, Jose I Neves, Paul E Grundy, Helena Brentani, Fernando A Soares, Beatriz De Camargo, and Dirce M Carraro. "Gene expression analysis of blastemal component reveals genes associated with relapse mechanism in Wilms tumour." In: *European Journal of Cancer* 47.18 (2011), pp. 2715–2722.

[339] Georgina N Masoud and Wei Li. "HIF-1α pathway: role, regulation and intervention for cancer therapy." In: *Acta Pharmaceutica Sinica B* 5.5 (2015), pp. 378–389.

[340] Glenn A Maston, Sara K Evans, and Michael R Green. "Transcriptional regulatory elements in the human genome." In: *Annu. Rev. Genomics Hum. Genet.* 7 (2006), pp. 29–59.

[341] Kenkichi Masutomi, Y Yu Evan, Shilagardy Khurts, Ittai Ben-Porath, Jennifer L Currier, Geoffrey B Metz, Mary W Brooks, Shuichi Kaneko, Seishi Murakami, James A DeCaprio, et al. "Telomerase maintains telomere structure in normal human cells." In: *Cell* 114.2 (2003), pp. 241–253.

[342] Vea Matys, Ellen Fricke, Robert Geffers, Ellen Gößling, Martin Haubrock, Reinhard Hehl, Klaus Hornischer, Dagmar Karas, Alexander E Kel, Olga V Kel-Margoulis, et al. "TRANSFAC®: transcriptional regulation, from patterns to profiles." In: *Nucleic acids research* 31.1 (2003), pp. 374–378.

[343] Paolo Mazzarello. "A unifying concept: the history of cell theory." In: *Nature cell biology* 1.1 (1999), E13–E15.

[344] DJ McCulley and BL Black. *Transcription factor pathways and congenital heart disease In: Heart Development: Current Topics in Developmental Biology*. 2012.

[345] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." In: *arXiv preprint arXiv:1802.03426* (2018).

[346] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. "The ensembl variant effect predictor." In: *Genome biology* 17.1 (2016), pp. 1–14.

[347] Robert C McLeay, Tom Lesluyes, Gabriel Cuellar Partida, and Timothy L Bailey. "Genome-wide in silico prediction of gene expression." In: *Bioinformatics* 28.21 (2012), pp. 2789–2796.

[348] European Society for Medical Oncology. *ESMO Clinical Practice Guidelines [Website]*. URL: https://www.esmo.org/guidelines (visited on 04/30/2021).

[349] FOUNDATION MEDICINE. *FoundationOne CDx*. 2021. URL: https://www.foundationmedicine.com/test/foundationone-cdx (visited on 11/25/2021).

[350] Yulia A Medvedeva, Andreas Lennartsson, Rezvan Ehsani, Ivan V Kulakovskiy, Ilya E Vorontsov, Pouda Panahandeh, Grigory Khimulya, Takeya Kasukawa, Finn Drabløs, FANTOM Consortium, et al. "EpiFactors: a comprehensive database of human epigenetic factors and complexes." In: *Database* 2015 (2015).

[351] Rohit Mehra, Sooryanarayana Varambally, Lei Ding, Ronglai Shen, Michael S Sabel, Debashis Ghosh, Arul M Chinnaiyan, and Celina G Kleer. "Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis." In: *Cancer research* 65.24 (2005), pp. 11259–11264.

[352] Rutika J Mehta, Rohit K Jain, Samuel Leung, Jennifer Choo, Torsten Nielsen, David Huntsman, Harikrishna Nakshatri, and Sunil Badve. "FOXA1 is an independent prognostic marker for ER-positive breast cancer." In: *Breast cancer research and treatment* 131.3 (2012), pp. 881–890.

[353] Jason H Mendler, Kati Maharry, Michael D Radmacher, Krzysztof Mrózek, Heiko Becker, Klaus H Metzeler, Sebastian Schwind, Susan P Whitman, Jihane Khalife, Jessica Kohlschmidt, et al. "RUNX1 mutations are associated with poor outcome in younger and older patients with cytogenetically normal acute myeloid leukemia and with distinct gene and MicroRNA expression signatures." In: *Journal of clinical oncology* 30.25 (2012), p. 3109.

[354] Nicolas S Merle, Sarah Elizabeth Church, Veronique Fremeaux-Bacchi, and Lubka T Roumenina. "Complement system part I–molecular mechanisms of activation and regulation." In: *Frontiers in immunology* 6 (2015), p. 262.

[355] Nicolas S Merle, Remi Noe, Lise Halbwachs-Mecarelli, Veronique Fremeaux-Bacchi, and Lubka T Roumenina. "Complement system part II: role in immunity." In: *Frontiers in immunology* 6 (2015), p. 257.

[356] Sally Metsuyanim, Naomi Pode-Shakked, Kai M Schmidt-Ott, Gilmor Keshet, Gideon Rechavi, Danith Blumental, and Benjamin Dekel. "Accumulation of malignant renal stem cells is associated with epigenetic changes in normal renal progenitor genes." In: *Stem Cells* 26.7 (2008), pp. 1808–1817.

[357] E Meyer, GV Aglyamova, and MV Matz. "Profiling gene expression responses of coral larvae (Acropora millepora) to elevated temperature and settlement inducers using a

novel RNA-Seq procedure." In: *Molecular ecology* 20.17 (2011), pp. 3599–3616.

[358] Huaiyu Mi, Betty Lazareva-Ulitsky, Rozina Loo, Anish Kejariwal, Jody Vandergriff, Steven Rabkin, Nan Guo, Anushya Muruganujan, Olivier Doremieux, Michael J Campbell, et al. "The PANTHER database of protein families, subfamilies, functions and pathways." In: *Nucleic acids research* 33.suppl_1 (2005), pp. D284–D288.

[359] James S Michaelson, Babara L Smith, Devon Bush, Bradford Diephuis, Leopoldo Fernandez, Nakul Valsangkar, and Jerry Younger. *CancerMath.net*. 2021. URL: http://www.lifemath.net/cancer/breastcancer/therapy/index.php (visited on 11/25/2021).

[360] Robert J Milner and J Gregor Sutcliffe. "Gene expression in rat brain." In: *Nucleic acids research* 11.16 (1983), pp. 5497–5520.

[361] Wang Ji Ming, L Bersani, and AJTJoI Mantovani. "Tumor necrosis factor is chemotactic for monocytes and polymorphonuclear leukocytes." In: *The Journal of Immunology* 138.5 (1987), pp. 1469–1474.

[362] Seyed M Moghadas, Meagan C Fitzpatrick, Pratha Sah, Abhishek Pandey, Affan Shoukat, Burton H Singer, and Alison P Galvani. "The implications of silent transmission for the control of COVID-19 outbreaks." In: *Proceedings of the National Academy of Sciences* 117.30 (2020), pp. 17513–17515.

[363] Reza Bayat Mokhtari, Tina S Homayouni, Narges Baluch, Evgeniya Morgatskaya, Sushil Kumar, Bikul Das, and Herman Yeger. "Combination therapy in combating cancer." In: *Oncotarget* 8.23 (2017), p. 38022.

[364] Pamela Moll, Michael Ante, Alexander Seitz, and Torsten Reda. "QuantSeq 3´ mRNA sequencing for RNA quantification." In: *Nature methods* 11.12 (2014), pp. i–iii.

[365] Daniel T Montoro, Adam L Haber, Moshe Biton, Vladimir Vinarsky, Brian Lin, Susan E Birket, Feng Yuan, Sijia Chen, Hui Min Leung, Jorge Villoria, et al. "A revised airway epithelial hierarchy includes CFTR-expressing ionocytes." In: *Nature* 560.7718 (2018), pp. 319–324.

[366] Jill E Moore, Michael J Purcaro, Henry E Pratt, Charles B Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A Davis, Alexander Dobin, Rajinder Kaul, et al. "Expanded encyclopaedias of DNA elements in the human and mouse genomes." In: *Nature* 583.7818 (2020), pp. 699–710.

[367] Lisa D Moore, Thuc Le, and Guoping Fan. "DNA methylation and its basic function." In: *Neuropsychopharmacology* 38.1 (2013), pp. 23–38.

[368] Daniel Morgensztern and Howard L McLeod. "PI3K/Akt/mTOR pathway as a target for cancer therapy." In: *Anti-cancer drugs* 16.8 (2005), pp. 797–803.

[369] Seiichi Mori, Rachel E Rempel, Jeffrey T Chang, Guang Yao, Anand S Lagoo, Anil Potti, Andrea Bild, and Joseph R Nevins. "Utilization of pathway signatures to reveal distinct types of B lymphoma in the Eμ-myc model and human diffuse large B-cell lymphoma." In: *Cancer research* 68.20 (2008), pp. 8525–8534.

[370] Nadya Morozova, Andrei Zinovyev, Nora Nonne, Linda-Louise Pritchard, Alexander N Gorban, and Annick Harel-Bellan. "Kinetic signatures of microRNA modes of action." In: *Rna* 18.9 (2012), pp. 1635–1655.

[371] Fabian Müller, Michael Scherer, Yassen Assenov, Pavlo Lutsik, Jörn Walter, Thomas Lengauer, and Christoph Bock. "RnBeads 2.0: comprehensive analysis of DNA methylation data." In: *Genome biology* 20.1 (2019), pp. 1–12.

[372] Mitsuyoshi Murata, Hiromi Nishiyori-Sueki, Miki Kojima-Ishiyama, Piero Carninci, Yoshihide Hayashizaki, and Masayoshi Itoh. "Detecting expressed genes using CAGE." In: *Transcription Factor Regulatory Networks*. Springer, 2014, pp. 67–85.

[373] Kenneth Murphy and Casey Weaver. *Janeway Immunologie*. Springer-Verlag, 2018.

[374] Haroon Naeem, Ralf Zimmer, Pegah Tavakkolkhah, and Robert Küffner. "Rigorous assessment of gene set enrichment tests." In: *Bioinformatics* 28.11 (2012), pp. 1480–1486.

[375] VP Nagraj, Neal E Magee, and Nathan C Sheffield. "LOLAweb: a containerized web server for interactive genomic locus overlap enrichment analysis." In: *Nucleic acids research* 46.W1 (2018), W194–W199.

[376] Dougu Nam and Seon-Young Kim. "Gene-set approach for expression pattern analysis." In: *Briefings in bioinformatics* 9.3 (2008), pp. 189–197.

[377] Anirudh Natarajan, Galip Gürkan Yardımcı, Nathan C Sheffield, Gregory E Crawford, and Uwe Ohler. "Predicting cell-type–specific gene expression from regions of open chromatin." In: *Genome research* 22.9 (2012), pp. 1711–1722.

[378] NCBI. *MAX MYC associated factor X [Homo sapiens (human)]*. 2021. URL: https://www.ncbi.nlm.nih.gov/gene/4149 (visited on 11/09/2021).

[379] NCBI. *NR2F2 nuclear receptor subfamily 2 group F member 2*. 2021. URL: https://www.ncbi.nlm.nih.gov/gene/7026 (visited on 11/08/2021).

[380] Daniel W Nebert. "Transcription factors and cancer: an overview." In: *Toxicology* 181 (2002), pp. 131–141.

[381] Daniel W Neef, Alex M Jaeger, and Dennis J Thiele. "Heat shock transcription factor 1 as a therapeutic target in neurodegenerative diseases." In: *Nature reviews Drug discovery* 10.12 (2011), pp. 930–944.

[382] Cancer Genome Atlas Network et al. "Comprehensive molecular portraits of human breast tumours." In: *Nature* 490.7418 (2012), p. 61.

[383] Cancer Genome Atlas Research Network. "Comprehensive molecular characterization of papillary renal-cell carcinoma." In: *New England Journal of Medicine* 374.2 (2016), pp. 135–145.

[384] Cancer Genome Atlas Research Network et al. "Comprehensive genomic characterization of squamous cell lung cancers." In: *Nature* 489.7417 (2012), p. 519.

[385] Richard M Neve, Koei Chin, Jane Fridlyand, Jennifer Yeh, Frederick L Baehner, Tea Fevr, Laura Clark, Nora Bayani, Jean-Philippe Coppe, Frances Tong, et al. "A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes." In: *Cancer cell* 10.6 (2006), pp. 515–527.

[386] Jerzy Neyman and Egon Sharpe Pearson. "IX. On the problem of the most efficient tests of statistical hypotheses." In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231.694-706 (1933), pp. 289–337.

[387] Frank Nielsen. "Hierarchical clustering." In: *Introduction to HPC with MPI for Data Science*. Springer, 2016, pp. 195–211.

[388] Sierra S Nishizaki, Natalie Ng, Shengcheng Dong, Robert S Porter, Cody Morterud, Colten Williams, Courtney Asman, Jessica A Switzenberg, and Alan P Boyle. "Predicting the effects of SNPs on transcription factor binding affinity." In: *Bioinformatics* 36.2 (2020), pp. 364–372.

[389] Joakim Nivre and Chiao-Ting Fang. "Universal dependency evaluation." In: *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*. 2017, pp. 86–95.

[390] Ryusuke Niwa and Frank J Slack. "The evolution of animal microRNA function." In: *Current opinion in genetics & development* 17.2 (2007), pp. 145–150.

[391] Trista E North, Terryl Stacy, Christina J Matheny, Nancy A Speck, and Marella FTR de Bruijn. "Runx1 is expressed in adult mouse hematopoietic stem cells and differentiating myeloid and lymphoid cells, but not in maturing erythroid cells." In: *Stem cells* 22.2 (2004), pp. 158–168.

[392] Jacob O'Brien, Heyam Hayder, Yara Zayed, and Chun Peng. "Overview of microRNA biogenesis, mechanisms of actions, and circulation." In: *Frontiers in endocrinology* 9 (2018), p. 402.

[393] German Federal Statistical Office. *A quarter of all deaths in Germany caused by cancer in 2019*. URL: https://www.destatis.de/EN/Press/2021/02/PE21_N010_231.html (visited on 03/11/2022).

[394] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. "KEGG: Kyoto encyclopedia of genes and genomes." In: *Nucleic acids research* 27.1 (1999), pp. 29–34.

[395] Chimari Okada, Eiki Yamashita, Soo Jae Lee, Satoshi Shibata, Jun Katahira, Atsushi Nakagawa, Yoshihiro Yoneda, and Tomitake Tsukihara. "A high-resolution structure of the pre-microRNA nuclear export machinery." In: *Science* 326.5957 (2009), pp. 1275–1279.

[396] Shinya Oki, Tazro Ohta, Go Shioi, Hideki Hatanaka, Osamu Ogasawara, Yoshihiro Okuda, Hideya Kawaji, Ryo Nakaki, Jun Sese, and Chikara Meno. "ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data." In: *EMBO reports* 19.12 (2018), e46255.

[397] Adam B Olshen, ES Venkatraman, Robert Lucito, and Michael Wigler. "Circular binary segmentation for the analysis of array-based DNA copy number data." In: *Biostatistics* 5.4 (2004), pp. 557–572.

[398] Rainer Opgen-Rhein and Korbinian Strimmer. "Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach." In: *Statistical Applications in Genetics & Molecular Biology* 6.1 (2007).

[399] Daniel P Oran and Eric J Topol. "Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review." In: *Annals of internal medicine* 173.5 (2020), pp. 362–367.

[400] Patrick A Ott, Zhuting Hu, Derin B Keskin, Sachet A Shukla, Jing Sun, David J Bozym, Wandi Zhang, Adrienne Luoma, Anita Giobbie-Hurder, Lauren Peter, et al. "An immunogenic personal neoantigen vaccine for patients with melanoma." In: *Nature* 547.7662 (2017), pp. 217–221.

[401] Maeve O'Huallachain, Felice-Alessio Bava, Mary Shen, Carolina Dallett, Sri Paladugu, Nikolay Samusik, Simon Yu, Razika Hussein, Grantland R Hillman, Samual Higgins, et al. "Ultra-high throughput single-cell analysis of proteins and RNAs by split-pool synthesis." In: *Communications biology* 3.1 (2020), pp. 1–19.

[402] PacBio. *PacBio Sequel Systems*. 2020. URL: https://www.pacb.com/products-and-services/sequel-system/ (visited on 12/26/2020).

[403] Soonmyung Paik, Steven Shak, Gong Tang, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L Baehner, Michael G Walker, Drew Watson, Taesung Park, et al. "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer." In: *New England Journal of Medicine* 351.27 (2004), pp. 2817–2826.

[404] Kostas A Papavassiliou and Athanasios G Papavassiliou. "Transcription factor drug targets." In: *Journal of cellular biochemistry* 117.12 (2016), pp. 2693–2696.

[405] Peter J Park. "ChIP–seq: advantages and challenges of a maturing technology." In: *Nature reviews genetics* 10.10 (2009), pp. 669–680.

[406] Debasish Paul, Venkatesh Chanukuppa, Panga Jaipal Reddy, Khushman Taunk, Ragini Adhav, Sanjeeva Srivastava, Manas Kumar Santra, and Srikanth Rapole. "Global proteomic profiling identifies etoposide chemoresistance markers in non-small cell lung carcinoma." In: *Journal of proteomics* 138 (2016), pp. 95–105.

[407] Linus Pauling and Robert B Corey. "Atomic coordinates and structure factors for two helical configurations of polypeptide chains." In: *Proceedings of the national academy of sciences of the United States of America* 37.5 (1951), p. 235.

[408] Karl Pearson. "VII. Note on regression and inheritance in the case of two parents." In: *proceedings of the royal society of London* 58.347-352 (1895), pp. 240–242.

[409] Mark Peplow. "The 100 000 genomes project." In: *Bmj* 353 (2016).

[410] Pavel A Pevzner, Haixu Tang, and Michael S Waterman. "An Eulerian path approach to DNA fragment assembly." In: *Proceedings of the national academy of sciences* 98.17 (2001), pp. 9748–9753.

[411] Theresa Phillips et al. "The role of methylation in gene expression." In: *Nature Education* 1.1 (2008), p. 116.

[412] Roger Pique-Regi, Jacob F Degner, Athma A Pai, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. "Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data." In: *Genome research* 21.3 (2011), pp. 447–455.

[413] Plotly. *Plotly*. URL: https://plotly.com (visited on 03/11/2021).

[414] Naomi Pode-Shakked, Rachel Shukrun, Michal Mark-Danieli, Peter Tsvetkov, Sarit Bahar, Sara Pri-Chen, Ronald S Goldstein, Eithan Rom-Gross, Yoram Mor, Edward Fridman, et al. "The isolation and characterization of renal cancer initiating cells from human Wilms' tumour xenografts unveils new therapeutic targets." In: *EMBO molecular medicine* 5.1 (2013), pp. 18–37.

[415] Fredrik Pontén, Karin Jirström, and Matthias Uhlen. "The Human Protein Atlas—a tool for pathology." In: *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 216.4 (2008), pp. 387–393.

[416] Dorin-Mirel Popescu, Rachel A Botting, Emily Stephenson, Kile Green, Simone Webb, Laura Jardine, Emily F Calderbank, Krzysztof Polanski, Issac Goh, Mirjana Efremova, et al. "Decoding human fetal liver haematopoiesis." In: *Nature* 574.7778 (2019), pp. 365–371.

[417] Camillo Porta, Chiara Paglino, and Alessandra Mosca. "Targeting PI3K/Akt/mTOR signaling in cancer." In: *Frontiers in oncology* 4 (2014), p. 64.

[418] Martin Preusse, Fabian J Theis, and Nikola S Mueller. "miTALOS v2: analyzing tissue specific microRNA function." In: *PLoS One* 11.3 (2016), e0151771.

[419] Kim D Pruitt, Jennifer Harrow, Rachel A Harte, Craig Wallin, Mark Diekhans, Donna R Maglott, Steve Searle, Catherine M Farrell, Jane E Loveland, Barbara J Ruef, et al. "The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes." In: *Genome research* 19.7 (2009), pp. 1316–1323.

[420] Kim D Pruitt, Tatiana Tatusova, Garth R Brown, and Donna R Maglott. "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy." In: *Nucleic acids research* 40.D1 (2012), pp. D130–D135.

[421] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." In: *Nucleic acids research* 35.suppl_1 (2007), pp. D61–D65.

[422] Sidharth V Puram, Itay Tirosh, Anuraag S Parikh, Anoop P Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina L Luo, Edmund A Mroz, Kevin S Emerick, et al. "Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer." In: *Cell* 171.7 (2017), pp. 1611–1624.

[423] Scott D Putney, Walter C Herlihy, and Paul Schimmel. "A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing." In: *Nature* 302.5910 (1983), pp. 718–721.

[424] Jun Qin, Xinpu Chen, Xin Xie, Ming-Jer Tsai, and Sophia Y Tsai. "COUP-TFII regulates tumor growth and metastasis by modulating tumor angiogenesis." In: *Proceedings of the National Academy of Sciences* 107.8 (2010), pp. 3687–3692.

[425] Peng Qiu. "Embracing the dropouts in single-cell RNA-seq analysis." In: *Nature communications* 11.1 (2020), pp. 1–9.

[426] Aaron R Quinlan. "BEDTools: the Swiss-army tool for genome feature analysis." In: *Current protocols in bioinformatics* 47.1 (2014), pp. 11–12.

[427] Dave Raggett, Arnaud Le Hors, Ian Jacobs, et al. "HTML 4.01 Specification." In: *W3C recommendation* 24 (1999).

[428] Bernard H Ramsahoye, Detlev Biniszkiewicz, Frank Lyko, Victoria Clark, Adrian P Bird, and Rudolf Jaenisch. "Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a." In: *Proceedings of the National Academy of Sciences* 97.10 (2000), pp. 5237–5242.

[429] Jon Ramsey, Kelly Butnor, Zhihua Peng, Tim Leclair, Jos van der Velden, Gary Stein, Jane Lian, and C Matthew Kinsey. "Loss of RUNX1 is associated with aggressive lung adenocarcinomas." In: *Journal of cellular physiology* 233.4 (2018), pp. 3487–3497.

[430] Ravi D Rao, Jan C Buckner, and Jann N Sarkaria. "Mammalian target of rapamycin (mTOR) inhibitors as anti-cancer agents." In: *Current Cancer Drug Targets* 4.8 (2004), pp. 621–635.

[431] David Ratel, Jean-Luc Ravanat, François Berger, and Didier Wion. "N6-methyladenine: the other methylated base of DNA." In: *Bioessays* 28.3 (2006), pp. 309–315.

[432] Timothy Ravasi, Christine Wells, Alistair Forest, David M Underhill, Brandon J Wainwright, Alan Aderem, Sean Grimmond, and David A Hume. "Generation of diversity in the innate immune system: macrophage heterogeneity arises from gene-autonomous transcriptional probability of individual inducible genes." In: *The Journal of Immunology* 168.1 (2002), pp. 44–50.

[433] RayBiotech. *Protein Arrays*. 2022. URL: https://www.raybiotech.com/protein-array (visited on 07/17/2022).

[434] Rachel E Rempel, Seiichi Mori, Maura Gasparetto, Michele A Glozak, Eran R Andrechek, Steven B Adler, Nina M Laakso, Anand S Lagoo, Robert Storms, Clay Smith, et al. "A role for E2F activities in determining the fate of Myc-induced lymphomagenesis." In: *PLoS genetics* 5.9 (2009), e1000640.

[435] Eric Rescorla et al. *HTTP Over TLS*. 2000.

[436] Jason A Reuter, Damek V Spacek, and Michael P Snyder. "High-throughput sequencing technologies." In: *Molecular cell* 58.4 (2015), pp. 586–597.

[437] Antonio Reverter, Nicholas J Hudson, Shivashankar H Nagaraj, Miguel Pérez-Enciso, and Brian P Dalrymple. "Regulatory impact factors: unraveling the transcriptional regulation of complex traits from expression data." In: *Bioinformatics* 26.7 (2010), pp. 896–904.

[438] Jean-Jack M Riethoven. "Regulatory regions in DNA: promoters, enhancers, silencers, and insulators." In: *Computational biology of transcription factor binding*. Springer, 2010, pp. 33–42.

[439] Alistair Ring and Mitch Dowsett. "Mechanisms of tamoxifen resistance." In: *Endocrine-related cancer* 11.4 (2004), pp. 643–658.

[440] Horst Rinne. "Taschenbuch der Statistik." In: (1997).

[441] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. "Normalization of RNA-seq data using factor analysis of control genes or samples." In: *Nature biotechnology* 32.9 (2014), pp. 896–902.

[442] Miguel N Rivera and Daniel A Haber. "Wilms' tumour: connecting tumorigenesis and organ development in the kidney." In: *Nature Reviews Cancer* 5.9 (2005), pp. 699–712.

[443] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." In: *Bioinformatics* 26.1 (2010), pp. 139–140.

[444] Mark D Robinson and Alicia Oshlack. "A scaling normalization method for differential expression analysis of RNA-seq data." In: *Genome biology* 11.3 (2010), pp. 1–9.

[445] Joëlle Roche. *The epithelial-to-mesenchymal transition in cancer*. 2018.

[446] Vijay K Rohatgi and AK Md Ehsanes Saleh. *An introduction to probability and statistics*. John Wiley & Sons, 2015.

[447] Helge G Roider, Aditi Kanhere, Thomas Manke, and Martin Vingron. "Predicting transcription factor affinities to DNA from a biophysical model." In: *Bioinformatics* 23.2 (2007), pp. 134–141.

[448] Helge G Roider, Thomas Manke, Sean O'Keeffe, Martin Vingron, and Stefan A Haas. "PASTAA: identifying transcription factors associated with sets of co-regulated genes." In: *Bioinformatics* 25.4 (2009), pp. 435–442.

[449] Alessandro Rosa and Ali H Brivanlou. "A regulatory circuitry comprised of miR-302 and the transcription factors OCT4 and NR2F2 regulates human embryonic stem cell differentiation." In: *The EMBO journal* 30.2 (2011), pp. 237–248.

[450] Assaf Rotem, Oren Ram, Noam Shoresh, Ralph A Sperling, Alon Goren, David A Weitz, and Bradley E Bernstein. "Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state." In: *Nature biotechnology* 33.11 (2015), pp. 1165–1172.

[451] Sugata Roy, Sebastian Schmeier, Erik Arner, Tanvir Alam, Suraj P Parihar, Mumin Ozturk, Ousman Tamgue, Hideya Kawaji, Michiel JL de Hoon, Masayoshi Itoh, et al. "Redefining the transcriptional regulatory dynamics of classically and alternatively activated macrophages by deepCAGE transcriptomics." In: *Nucleic acids research* 43.14 (2015), pp. 6969–6982.

[452] Sameek Roychowdhury, Matthew K Iyer, Dan R Robinson, Robert J Lonigro, Yi-Mi Wu, Xuhong Cao, Shanker Kalyana-Sundaram, Lee Sam, O Alejandro Balbin, Michael J Quist, et al. "Personalized oncology through integrative high-throughput sequencing: a pilot study." In: *Science translational medicine* 3.111 (2011), 111ra121–111ra121.

[453] J Graham Ruby, Calvin H Jan, and David P Bartel. "Intronic microRNA precursors that bypass Drosha processing." In: *Nature* 448.7149 (2007), pp. 83–86.

[454] Andreas Ruepp, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Michael Stransky, Brigitte Waegele, Thorsten Schmidt, Octave Noubibou Doudieu, Volker Stümpflen, et al. "CORUM: the comprehensive resource of mammalian protein complexes." In: *Nucleic acids research* 36.suppl_1 (2007), pp. D646–D650.

[455] Annemieke JM de Ruijter, Albert H van GENNIP, Huib N Caron, Stephan Kemp, and André BP van KUILENBURG. "Histone deacetylases (HDACs): characterization of the classical HDAC family." In: *Biochemical Journal* 370.3 (2003), pp. 737–749.

[456] Katja Rust, Laurean E Byrnes, Kevin Shengyang Yu, Jason S Park, Julie B Sneddon, Aaron D Tward, and Todd G Nystul. "A single-cell atlas and lineage analysis of the adult Drosophila ovary." In: *Nature communications* 11.1 (2020), pp. 1–17.

[457] Michael Sachs, Courtney Onodera, Kathryn Blaschke, Kevin T Ebata, Jun S Song, and Miguel Ramalho-Santos. "Bivalent chromatin marks developmental regulatory genes in the mouse embryonic germline in vivo." In: *Cell reports* 3.6 (2013), pp. 1777–1784.

[458] Anjanabha Saha, Jacqueline Wittmeyer, and Bradley R Cairns. "Chromatin remodelling: the industrial revolution of DNA around histones." In: *Nature reviews Molecular cell biology* 7.6 (2006), pp. 437–447.

[459] Ugur Sahin, Evelyna Derhovanessian, Matthias Miller, Björn-Philipp Kloke, Petra Simon, Martin Löwer, Valesca Bukur, Arbel D Tadmor, Ulrich Luxemburger, Barbara Schrörs, et al. "Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer." In: *Nature* 547.7662 (2017), pp. 222–226.

[460] Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. "Single-cell RNA-seq: advances and future challenges." In: *Nucleic acids research* 42.14 (2014), pp. 8845–8860.

[461] Joseph Sambrook and David W Russell. "Fragmentation of DNA by sonication." In: *Cold spring harbor protocols* 2006.4 (2006), pdb–prot4538.

[462] Francisco Sanchez-Vega, Marco Mina, Joshua Armenia, Walid K Chatila, Augustin Luna, Konnor C La, Sofia Dimitriadoy, David L Liu, Havish S Kantheti, Sadegh Saghafinia, et al. "Oncogenic signaling pathways in the cancer genome atlas." In: *Cell* 173.2 (2018), pp. 321–337.

[463] F Sanger, Ar R Coulson, GF Hong, DF Hill, and GB d Petersen. "Nucleotide sequence of bacteriophage λ DNA." In: *Journal of molecular biology* 162.4 (1982), pp. 729–773.

[464] Frederick Sanger, Gilian M Air, Bart G Barrell, Nigel L Brown, Alan R Coulson, John C Fiddes, CA Hutchison, Patrick M Slocombe, and Mo Smith. "Nucleotide sequence of bacteriophage φX174 DNA." In: *nature* 265.5596 (1977), pp. 687–695.

[465] Frederick Sanger, Steven Nicklen, and Alan R Coulson. "DNA sequencing with chain-terminating inhibitors." In: *Proceedings of the national academy of sciences* 74.12 (1977), pp. 5463–5467.

[466] Steffen Sass, Florian Buettner, Nikola S Mueller, and Fabian J Theis. "RAMONA: a Web application for gene set analysis on multilevel omics data." In: *Bioinformatics* 31.1 (2015), pp. 128–130.

[467] Franklin E Satterthwaite. "Synthesis of variance." In: *Psychometrika* 6.5 (1941), pp. 309–316.

[468] Franklin E Satterthwaite. "An approximate distribution of estimates of variance components." In: *Biometrics bulletin* 2.6 (1946), pp. 110–114.

[469] Boris Schäling. *The boost C++ libraries*. Boris Schäling, 2011.

[470] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." In: *Science* 270.5235 (1995), pp. 467–470.

[471] MJ Schleiden. "Arch Anat Physiol." In: *Wiss Med* 13 (1838), pp. 137–176.

[472] Florian Schmidt, Nina Gasparoni, Gilles Gasparoni, Kathrin Gianmoena, Cristina Cadenas, Julia K Polansky, Peter Ebert, Karl Nordström, Matthias Barann, Anupam Sinha, et al. "Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction." In: *Nucleic acids research* 45.1 (2017), pp. 54–66.

[473] Florian Schmidt, Fabian Kern, Peter Ebert, Nina Baumgarten, and Marcel H Schulz. "TEPIC 2—an extended framework for transcription factor binding prediction and integrative epigenomic analysis." In: *Bioinformatics* 35.9 (2019), pp. 1608–1609.

[474] Lara Schneider, Daniel Stöckel, **Tim Kehl**, Andreas Gerasch, Nicole Ludwig, Petra Leidinger, Hanno Huwer, Stefan Tenzer, Oliver Kohlbacher, Andreas Hildebrandt, et al. "DrugTargetInspector: An assistance tool for patient treatment stratification." In: *International journal of cancer* 138.7 (2016), pp. 1765–1776.

[475] Lara Schneider, **Tim Kehl**, Kristina Thedinga, Nadja Liddy Grammes, Christina Backes, Christopher Mohr, Benjamin Schubert, Kerstin Lenhof, Nico Gerstner, Andreas Daniel Hartkopf, et al. "ClinOmicsTrailbc: a visual analytics tool for breast cancer treatment stratification." In: *Bioinformatics* 35.24 (2019), pp. 5171–5181.

[476] Lara Kristina Schneider. "Multi-omics integrative analyses for decision support systems in personalized cancer treatment." In: (2020).

[477] Robert D Schreiber, Lloyd J Old, and Mark J Smyth. "Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion." In: *Science* 331.6024 (2011), pp. 1565–1570.

[478] Werner Schroth, Lydia Antoniadou, Peter Fritz, Matthias Schwab, Thomas Muerdter, Ulrich M Zanger, Wolfgang Simon, Michel Eichelbaum, and Hiltrud Brauch. "Breast cancer treatment outcome with adjuvant tamoxifen relative to patient

CYP2D6 and CYP2C19 genotypes." In: *Journal of Clinical Oncology* 25.33 (2007), pp. 5187–5193.

[479] Benjamin Schubert, Luis de la Garza, Christopher Mohr, Mathias Walzer, and Oliver Kohlbacher. "ImmunoNodes–graphical development of complex immunoinformatics workflows." In: *BMC bioinformatics* 18.1 (2017), pp. 1–7.

[480] Anja Schuetz, Didier Nana, Charlotte Rose, Georg Zocher, Maja Milanovic, Jessica Koenigsmann, Rosel Blasig, Udo Heinemann, and Dirk Carstanjen. "The structure of the Klf4 DNA-binding domain links to self-renewal and macrophage differentiation." In: *Cellular and Molecular Life Sciences* 68.18 (2011), pp. 3121–3131.

[481] Theodor Schwann and Friedrich Hünseler. *Mikroskopische Untersuchungen über die Ubereinstimmung in der Struktur und dem Wachstume der Tiere und Pflanzen*. 176. W. Engelmann, 1910.

[482] E Senkus, S Kyriakides, S Ohno, F Penault-Llorca, P Poortmans, E Rutgers, Sophia Zackrisson, and F Cardoso. "Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up." In: *Annals of oncology* 26 (2015), pp. v8–v30.

[483] Frances A Shepherd, Giuseppe Giaccone, Lesley Seymour, Channa Debruyne, Andrea Bezjak, Vera Hirsh, Michael Smylie, Sheldon Rubin, Heidi Martins, Alan Lamont, et al. "Prospective, randomized, double-blind, placebo-controlled trial of marimastat after response to first-line chemotherapy in patients with small-cell lung cancer: a trial of the National Cancer Institute of Canada-Clinical Trials Group and the European Organization for Research and Treatment of Cancer." In: *Journal of clinical oncology* 20.22 (2002), pp. 4434–4439.

[484] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. "dbSNP: the NCBI database of genetic variation." In: *Nucleic acids research* 29.1 (2001), pp. 308–311.

[485] Richard I Sherwood, Tatsunori Hashimoto, Charles W O'donnell, Sophia Lewis, Amira A Barkal, John Peter Van Hoff, Vivek Karun, Tommi Jaakkola, and David K Gifford. "Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape." In: *Nature biotechnology* 32.2 (2014), pp. 171–178.

[486] Rachel Shukrun, Naomi Pode-Shakked, Oren Pleniceanu, Dorit Omer, Einav Vax, Eyal Peer, Sara Pri-Chen, Jasmine Jacob, Qianghua Hu, Orit Harari-Steinberg, et al. "Wilms' tumor blastemal stem cells dedifferentiate to propagate the tumor bulk." In: *Stem Cell Reports* 3.1 (2014), pp. 24–33.

[487] Zbyněk Šidák. "Rectangular confidence regions for the means of multivariate normal distributions." In: *Journal of the American Statistical Association* 62.318 (1967), pp. 626–633.

[488] Hifzur Rahman Siddique and Mohammad Saleem. "Role of BMI1, a stem cell factor, in cancer recurrence and chemoresistance: preclinical and clinical evidences." In: *Stem cells* 30.3 (2012), pp. 372–378.

[489] Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, and Ahmedin Jemal. "Cancer statistics, 2021." In: *CA: a cancer journal for clinicians* 71.1 (2021), pp. 7–33.

[490] Graham W Slack and Randy D Gascoyne. "MYC and aggressive B-cell lymphomas." In: *Advances in anatomic pathology* 18.3 (2011), pp. 219–228.

[491] Elzbieta A Slodkowska and Jeffrey S Ross. "MammaPrint™ 70-gene signature: another milestone in personalized medical care for breast cancer patients." In: *Expert review of molecular diagnostics* 9.5 (2009), pp. 417–422.

[492] Geir Slupphaug, Bodil Kavli, and Hans E Krokan. "The interacting pathways for prevention and repair of oxidative DNA damage." In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 531.1-2 (2003), pp. 231–251.

[493] Michal Slyper, Amit Shahar, Anat Bar-Ziv, Roy Z Granit, Tamar Hamburger, Bella Maly, Tamar Peretz, and Ittai Ben-Porath. "Control of breast cancer growth and initiation by the stem cell–associated transcription factor TCF3." In: *Cancer research* 72.21 (2012), pp. 5613–5624.

[494] Stephen T Smale and James T Kadonaga. "The RNA polymerase II core promoter." In: *Annual review of biochemistry* 72.1 (2003), pp. 449–479.

[495] Marcel Smid, Robert RJ Coebergh van den Braak, Harmen JG van de Werken, Job van Riet, Anne van Galen, Vanja de Weerd, Michelle van der Vlugt-Daane, Sandra I Bril, Zarina S Lalmahomed, Wigard P Kloosterman, et al. "Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data performs similarly in intersample analyses while improving intrasample comparisons." In: *BMC bioinformatics* 19.1 (2018), pp. 1–13.

[496] Nickolay Smirnov. "Table for estimating the goodness of fit of empirical distributions." In: *The annals of mathematical statistics* 19.2 (1948), pp. 279–281.

[497] Mike L Smith, Keith A Baggerly, Henrik Bengtsson, Matthew E Ritchie, and Kasper D Hansen. "illuminaio: An open source IDAT parsing tool for Illumina microarrays." In: *F1000Research* 2 (2013).

[498] Gordon K Smyth. "Limma: linear models for microarray data." In: *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, 2005, pp. 397–420.

[499] SmartBear Software. *Swagger*. URL: https://swagger.io/ (visited on 03/11/2021).

[500] Lingyun Song and Gregory E Crawford. "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells." In: *Cold Spring Harbor Protocols* 2010.2 (2010), pdb–prot5384.

[501] Sarah Song and Michael A Black. "Microarray-based gene set analysis: a comparison of current methods." In: *BMC bioinformatics* 9.1 (2008), pp. 1–14.

[502] Eric de Sousa, Dário Ligeiro, Joana R Lérias, Chao Zhang, Chiara Agrati, Mohamed Osman, Sherif A El-Kafrawy, Esam I Azhar, Giuseppe Ippolito, Fu-Sheng Wang, et al. "Mortality in COVID-19 disease patients: Correlating the association of major histocompatibility complex (MHC) with severe acute respiratory syndrome 2 (SARS-CoV-2) variants." In: *International Journal of Infectious Diseases* 98 (2020), pp. 454–459.

[503] Julie Soutourina. "Transcription regulation by the Mediator complex." In: *Nature reviews Molecular cell biology* 19.4 (2018), pp. 262–274.

[504] Charles Spearman. "The proof and measurement of association between two things." In: (1961).

[505] Neil H Spencer. *Essentials of multivariate data analysis*. CRC press, 2013.

[506] Sriganesh Srihari, Murugan Kalimutho, Samir Lal, Jitin Singla, Dhaval Patel, Peter T Simpson, Kum Kum Khanna, and Mark A Ragan. "Understanding the functional impact of copy number alterations in breast cancer using a network modeling approach." In: *Molecular BioSystems* 12.3 (2016), pp. 963–972.

[507] Luisa Statello, Chun-Jie Guo, Ling-Ling Chen, and Maite Huarte. "Gene regulation by long non-coding RNAs and its biological functions." In: *Nature Reviews Molecular Cell Biology* 22.2 (2021), pp. 96–118.

[508] Lincoln Stein. *Generic Feature Format Version 3 (GFF3)*. 2020. URL: https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md (visited on 02/01/2022).

[509] Daniel Stöckel, Oliver Müller, **Tim Kehl**, Andreas Gerasch, Christina Backes, Alexander Rurainski, Andreas Keller, Michael Kaufmann, and Hans-Peter Lenhof. "NetworkTrail—a web service for identifying and visualizing deregulated subnetworks." In: *Bioinformatics* 29.13 (2013), pp. 1702–1703.

[510] Daniel Stöckel, **Tim Kehl**, Patrick Trampert, Lara Schneider, Christina Backes, Nicole Ludwig, Andreas Gerasch, Michael Kaufmann, Manfred Gessler, Norbert Graf, et al. "Multi-omics enrichment analysis using the GeneTrail2 web service." In: *Bioinformatics* 32.10 (2016), pp. 1502–1508.

[511] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. "Simultaneous epitope and transcriptome measurement in single cells." In: *Nature methods* 14.9 (2017), p. 865.

[512] Robert O Stuart, Kevin T Bush, and Sanjay K Nigam. "Changes in gene expression patterns in the ureteric bud and metanephric mesenchyme in models of kidney development." In: *Kidney international* 64.6 (2003), pp. 1997–2008.

[513] Aravind Subramanian, Heidi Kuehn, Joshua Gould, Pablo Tamayo, and Jill P Mesirov. "GSEA-P: a desktop application for Gene Set Enrichment Analysis." In: *Bioinformatics* 23.23 (2007), pp. 3251–3253.

[514] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550.

[515] Haoqi Sun, Haiping Wang, Ruixin Zhu, Kailin Tang, Qin Gong, Juan Cui, Zhiwei Cao, and Qi Liu. "iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis." In: *Bioinformatics* 30.5 (2014), pp. 737–739.

[516] Myong-Hee Sung, Songjoon Baek, and Gordon L Hager. "Genome-wide footprinting: ready for prime time?" In: *Nature methods* 13.3 (2016), pp. 222–228.

[517] Myong-Hee Sung, Michael J Guertin, Songjoon Baek, and Gordon L Hager. "DNase footprint signatures are dictated by factor dynamics and DNA sequence." In: *Molecular cell* 56.2 (2014), pp. 275–285.

[518] Valentine Svensson, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic, and Sarah A Teichmann. "Power analysis of single-cell RNA-sequencing experiments." In: *Nature methods* 14.4 (2017), pp. 381–387.

[519] Gábor J Székely and Maria L Rizzo. "Brownian distance covariance." In: *The annals of applied statistics* 3.4 (2009), pp. 1236–1265.

[520] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. "Measuring and testing dependence by correlation of distances." In: *The annals of statistics* 35.6 (2007), pp. 2769–2794.

[521] Bo Tang, Guangying Qi, Fang Tang, Shengguang Yuan, Zhenran Wang, Xingsi Liang, Bo Li, Shuiping Yu, Jie Liu, Qi Huang, et al. "JARID1B promotes metastasis and epithelial-mesenchymal transition via PTEN/AKT signaling in hepatocellular carcinoma cells." In: *Oncotarget* 6.14 (2015), p. 12723.

[522] Andrea Tanzer and Peter F Stadler. "Molecular evolution of a microRNA cluster." In: *Journal of molecular biology* 339.2 (2004), pp. 327–335.

[523] Bootstrap Development Team et al. *Bootstrap*. URL: https://getbootstrap.com/ (visited on 03/11/2021).

[524] The Thymeleaf Team. *Thymeleaf*. URL: https://www.thymeleaf.org/ (visited on 03/11/2021).

[525] Oxford Nanopore Technologies. *MinION*. 2020. URL: https://nanoporetech.com/products/minion (visited on 12/26/2020).

[526] Wee-Wei Tee and Danny Reinberg. "Chromatin features and the epigenetic regulation of pluripotency states in ESCs." In: *Development* 141.12 (2014), pp. 2376–2390.

[527] Sarah Teichmann and Mirjana Efremova. "Method of the Year 2019: single-cell multimodal omics." In: *Nat. Methods* 17.1 (2020).

[528] Markus F Templin, Dieter Stoll, Monika Schrenk, Petra C Traub, Christian F Vöhringer, and Thomas O Joos. "Protein microarray technology." In: *Drug discovery today* 7.15 (2002), pp. 815–822.

[529] **Tim Kehl**, Christina Backes, Fabian Kern, Tobias Fehlmann, Nicole Ludwig, Eckart Meese, Hans-Peter Lenhof, and Andreas Keller. "About miRNAs, miRNA seeds, target genes and target pathways." In: *Oncotarget* 8.63 (2017), p. 107167.

[530] **Tim Kehl**, Fabian Kern, Christina Backes, Tobias Fehlmann, Daniel Stöckel, Eckart Meese, Hans-Peter Lenhof, and Andreas Keller. "miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database." In: *Nucleic acids research* 48.D1 (2020), pp. D142–D147.

[531] **Tim Kehl**, Lara Schneider, Kathrin Kattler, Daniel Stöckel, Jenny Wegert, Nico Gerstner, Nicole Ludwig, Ute Distler, Markus Schick, Ulrich Keller, et al. "REGGAE: a novel approach for the identification of key transcriptional regulators." In: *Bioinformatics* 1 (2018), p. 8.

[532] **Tim Kehl**, Lara Schneider, Kathrin Kattler, Daniel Stöckel, Jenny Wegert, Nico Gerstner, Nicole Ludwig, Ute Distler, Stefan Tenzer, Manfred Gessler, et al. "The role of TCF3 as potential master regulator in blastemal Wilms tumors." In: *International journal of cancer* 144.6 (2019), pp. 1432–1443.

[533] **Tim Kehl**, Lara Schneider, Florian Schmidt, Daniel Stöckel, Nico Gerstner, Christina Backes, Eckart Meese, Andreas Keller, Marcel H Schulz, and Hans-Peter Lenhof. "RegulatorTrail: a web service for the identification of key transcriptional regulators." In: *Nucleic acids research* 45.W1 (2017), W146–W153.

[534] Reuben Thomas, Sean Thomas, Alisha K Holloway, and Katherine S Pollard. "Features that define the best ChIP-seq peak calling algorithms." In: *Briefings in bioinformatics* 18.3 (2017), pp. 441–450.

[535] Morgane Thomas-Chollier, Andrew Hufton, Matthias Heinig, Sean O'keeffe, Nassim El Masri, Helge G Roider, Thomas Manke, and Martin Vingron. "Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs." In: *Nature protocols* 6.12 (2011), pp. 1860–1869.

[536] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge." In: *Contemporary oncology* 19.1A (2015), A68.

[537] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. "From Louvain to Leiden: guaranteeing well-connected communities." In: *Scientific reports* 9.1 (2019), pp. 1–12.

[538] Tiina T Tuomisto, Mervi S Riekkinen, Helena Viita, Anna-Liisa Levonen, and Seppo Ylä-Herttuala. "Analysis of gene and protein expression during monocyte-macrophage differentiation and cholesterol loading—cDNA and protein array study." In: *Atherosclerosis* 180.2 (2005), pp. 283–291.

[539] Sean Turner. "Transport layer security." In: *IEEE Internet Computing* 18.6 (2014), pp. 60–63.

[540] Julianne D Twomey, Nina N Brahme, and Baolin Zhang. "Drug-biomarker co-development in oncology–20 years and counting." In: *Drug resistance updates* 30 (2017), pp. 48–62.

[541] Emil R Unanue. "Antigen-presenting function of the macrophage." In: *Annual review of immunology* 2.1 (1984), pp. 395–428.

[542] Anton Valouev, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data." In: *Nature methods* 5.9 (2008), pp. 829–834.

[543] Nynke L Van Berkum, Erez Lieberman-Aiden, Louise Williams, Maxim Imakaev, Andreas Gnirke, Leonid A Mirny, Job Dekker, and Eric S Lander. "Hi-C: a method to study the three-dimensional architecture of genomes." In: *JoVE (Journal of Visualized Experiments)* 39 (2010), e1869.

[544] Matthew G Vander Heiden, Lewis C Cantley, and Craig B Thompson. "Understanding the Warburg effect: the metabolic requirements of cell proliferation." In: *science* 324.5930 (2009), pp. 1029–1033.

[545] Laura J Van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin Van Der Kooy, Matthew J Marton, Anke T Witteveen, et al. "Gene expression profiling predicts clinical outcome of breast cancer." In: *nature* 415.6871 (2002), pp. 530–536.

[546] Juan M Vaquerizas, Sarah K Kummerfeld, Sarah A Teichmann, and Nicholas M Luscombe. "A census of human transcription factors: function, expression and evolution." In: *Nature Reviews Genetics* 10.4 (2009), pp. 252–263.

[547] Ales Varabyou, Steven L Salzberg, and Mihaela Pertea. "Effects of transcriptional noise on estimates of gene and transcript expression in RNA sequencing experiments." In: *Genome research* 31.2 (2021), pp. 301–308.

[548] Enrique I Velazquez-Villarreal, Shamoni Maheshwari, Jon Sorenson, Ian T Fiddes, Vijay Kumar, Yifeng Yin, Michelle G Webb, Claudia Catalanotti, Mira Grigorova, Paul A Edwards, et al. "Single-cell sequencing of genomic DNA resolves subclonal heterogeneity in a melanoma cell line." In: *Communications biology* 3.1 (2020), pp. 1–8.

[549] DL Van der Velden, CML Van Herpen, HWM Van Laarhoven, EF Smit, HJM Groen, Stefan Martin Willems, PM Nederlof, MHG Langenberg, Edwin Cuppen, Stefan Sleijfer, et al. "Molecular tumor boards: current practice and future needs." In: *Annals of Oncology* 28.12 (2017), pp. 3070–3075.

[550] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. "The sequence of the human genome." In: *science* 291.5507 (2001), pp. 1304–1351.

[551] Ioannis S Vlachos, Konstantinos Zagganas, Maria D Paraskevopoulou, Georgios Georgakilas, Dimitra Karagkouni, Thanasis Vergoulis, Theodore Dalamagas, and Artemis G Hatzigeorgiou. "DIANA-miRPath v3. 0: deciphering microRNA function with experimental support." In: *Nucleic acids research* 43.W1 (2015), W460–W466.

[552] World Wide Web Consortium (W3C) et al. URL: https://www.w3.org/Style/CSS/ (visited on 03/11/2021).

[553] World Wide Web Consortium (W3C) et al. *HTML Living Standard*. URL: https://html.spec.whatwg.org/ (visited on 03/11/2021).

[554] Günter P Wagner, Koryu Kin, and Vincent J Lynch. "Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples." In: *Theory in biosciences* 131.4 (2012), pp. 281–285.

[555] Fazli Wahid, Adeeb Shehzad, Taous Khan, and You Young Kim. "MicroRNAs: synthesis, mechanism, function, and recent clinical trials." In: *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1803.11 (2010), pp. 1231–1243.

[556] Colum P Walsh and Timothy H Bestor. "Cytosine methylation and mammalian development." In: *Genes & development* 13.1 (1999), pp. 26–34.

[557] Gang G Wang, C David Allis, and Ping Chi. "Chromatin remodeling and cancer, Part I: Covalent histone modifications." In: *Trends in molecular medicine* 13.9 (2007), pp. 363–372.

[558] Gang G Wang, C David Allis, and Ping Chi. "Chromatin remodeling and cancer, Part II: ATP-dependent chromatin remodeling." In: *Trends in molecular medicine* 13.9 (2007), pp. 373–380.

[559] Lixin Wang, Joan S Brugge, and Kevin A Janes. "Intersection of FOXO-and RUNX1-mediated gene expression programs in single breast epithelial cells during morphogenesis and tumor progression." In: *Proceedings of the National Academy of Sciences* 108.40 (2011), E803–E812.

[560] Ya-Wen Wang, Xu Chen, Ji-Wei Gao, Hui Zhang, Ran-Ran Ma, Zu-Hua Gao, and Peng Gao. "High expression of cAMP responsive element binding protein 1 (CREB1) is associated with metastasis, tumor stage and poor outcome in gastric cancer." In: *Oncotarget* 6.12 (2015), p. 10646.

[561] Zheng Wang, Efrat Oron, Brynna Nelson, Spiro Razis, and Natalia Ivanova. "Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells." In: *Cell stem cell* 10.4 (2012), pp. 440–454.

[562] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics." In: *Nature reviews genetics* 10.1 (2009), pp. 57–63.

[563] Otto Warburg. "On the origin of cancer cells." In: *Science* 123.3191 (1956), pp. 309–314.

[564] Joe H Ward Jr. "Hierarchical grouping to optimize an objective function." In: *Journal of the American statistical association* 58.301 (1963), pp. 236–244.

[565] Dennis K Watson, David P Turner, Melissa N Scheiber, Victoria J Findlay, and Patricia M Watson. "ETS transcription factor expression and conversion during prostate and breast cancer progression." In: *The Open Cancer Journal* 3.1 (2010).

[566] James D Watson, Francis HC Crick, et al. "A structure for deoxyribose nucleic acid." In: *Nature* 171.4356 (1953), pp. 737–738.

[567] Jenny Wegert, Sabrina Bausenwein, Sabine Roth, Norbert Graf, Eva Geissinger, and Manfred Gessler. "Characterization of primary Wilms tumor cultures as an in vitro model." In: *Genes, Chromosomes and Cancer* 51.1 (2012), pp. 92–104.

[568] Jenny Wegert, Naveed Ishaque, Romina Vardapour, Christina Geörg, Zuguang Gu, Matthias Bieg, Barbara Ziegler, Sabrina Bausenwein, Nasenien Nourkami, Nicole Ludwig, et al. "Mutations in the SIX1/2 pathway and the DROSHA/DGCR8 miRNA microprocessor complex underlie high-risk blastemal type Wilms tumors." In: *Cancer cell* 27.2 (2015), pp. 298–311.

[569] Robert A Weinberg. *The biology of cancer*. Garland science, 2013.

[570] A Weirich, I Leuschner, D Harms, GM Vujanic, J Tröger, U Abel, N Graf, D Schmidt, R Ludwig, and PA Voute. "Clinical impact of histologic subtypes in localized non-anaplastic nephroblastoma treated according to the trial and study SIOP-9/GPOH." In: *Annals of oncology* 12.3 (2001), pp. 311–319.

[571] Bernard L Welch. "The significance of the difference between two means when the population variances are unequal." In: *Biometrika* 29.3/4 (1938), pp. 350–362.

[572] Bernard L Welch. "The generalization of student's' problem when several different population variances are involved." In: *Biometrika* 34.1/2 (1947), pp. 28–35.

[573] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations." In: *Nucleic acids research* 42.D1 (2014), pp. D1001–D1006.

[574] Mike West, Carrie Blanchette, Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, John A Olson, Jeffrey R Marks, and Joseph R Nevins. "Predicting the clinical status of human breast cancer by using gene expression profiles." In: *Proceedings of the National Academy of Sciences* 98.20 (2001), pp. 11462–11467.

[575] Kris A. Wetterstrand. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. 2020. URL: www.genome.gov/sequencingcostsdata (visited on 10/07/2020).

[576] Kris A. Wetterstrand. *The Cost of Sequencing a Human Genome*. 2020. URL: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost (visited on 10/07/2020).

[577] George M Whitesides. "The origins and the future of microfluidics." In: *Nature* 442.7101 (2006), pp. 368–373.

[578] World Health Organization (WHO). *WHO Coronavirus (COVID-19) Dashboard*. 2021. URL: https://covid19.who.int/ (visited on 10/28/2021).

[579] Wikipedia. *Hamming-Abstand*. 2021. URL: https://de.wikipedia.org/wiki/Hamming-Abstand (visited on 09/27/2021).

[580] Wikipedia. *URL*. 2021. URL: https://en.wikipedia.org/wiki/URL (visited on 02/21/2021).

[581] Aaron J Wilk, Arjun Rustagi, Nancy Q Zhao, Jonasel Roque, Giovanny J Martínez-Colón, Julia L McKechnie, Geoffrey T Ivison, Thanmayi Ranganath, Rosemary Vergara, Taylor Hollis, et al. "A single-cell atlas of the peripheral immune response in patients with severe COVID-19." In: *Nature medicine* 26.7 (2020), pp. 1070–1076.

[582] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. "DrugBank 5.0: a major update to the DrugBank database for 2018." In: *Nucleic acids research* 46.D1 (2018), pp. D1074–D1082.

[583] Katherine Wolstencroft, Robert Haines, Donal Fellows, Alan Williams, David Withers, Stuart Owen, Stian Soiland-Reyes, Ian Dunlop, Aleksandra Nenadic, Paul Fisher, et al. "The Taverna workflow suite: designing and executing workflows of

Web Services on the desktop, web or in the cloud." In: *Nucleic acids research* 41.W1 (2013), W557–W561.

[584] Woodring E Wright, Mieczyslaw A Piatyszek, William E Rainey, William Byrd, and Jerry W Shay. "Telomerase activity in human germline and embryonic tissues and cells." In: *Developmental genetics* 18.2 (1996), pp. 173–179.

[585] Angela R Wu, Norma F Neff, Tomer Kalisky, Piero Dalerba, Barbara Treutlein, Michael E Rothenberg, Francis M Mburu, Gary L Mantalas, Sopheak Sim, Michael F Clarke, et al. "Quantitative assessment of single-cell RNA-sequencing methods." In: *Nature methods* 11.1 (2014), p. 41.

[586] Song Wu, Scott Powers, Wei Zhu, and Yusuf A Hannun. "Substantial contribution of extrinsic risk factors to cancer development." In: *Nature* 529.7584 (2016), pp. 43–47.

[587] Tao P Wu, Tao Wang, Matthew G Seetin, Yongquan Lai, Shijia Zhu, Kaixuan Lin, Yifei Liu, Stephanie D Byrum, Samuel G Mackintosh, Mei Zhong, et al. "DNA methylation on N 6-adenine in mammalian embryonic stem cells." In: *Nature* 532.7599 (2016), pp. 329–333.

[588] X Wu and Mick Watson. "CORNA: testing gene lists for regulation by microRNAs." In: *Bioinformatics* 25.6 (2009), pp. 832–833.

[589] Zunyou Wu and Jennifer M McGoogan. "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention." In: *Jama* 323.13 (2020), pp. 1239–1242.

[590] Ken Yamazaki, Yohei Masugi, Kathryn Effendi, Hanako Tsujikawa, Nobuyoshi Hiraoka, Minoru Kitago, Masahiro Shinoda, Osamu Itano, Minoru Tanabe, Yuko Kitagawa, et al. "Upregulated SMAD3 promotes epithelial–mesenchymal transition and predicts poor prognosis in pancreatic ductal adenocarcinoma." In: *Laboratory investigation* 94.6 (2014), pp. 683–691.

[591] Chuhu Yang, Eugene Bolotin, Tao Jiang, Frances M Sladek, and Ernest Martinez. "Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters." In: *Gene* 389.1 (2007), pp. 52–65.

[592] Jian-Hua Yang, Jun-Hao Li, Shan Jiang, Hui Zhou, and Liang-Hu Qu. "ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data." In: *Nucleic acids research* 41.D1 (2013), pp. D177–D187.

[593] Jing Yang, Hui Yu, Bao-Hong Liu, Zhongming Zhao, Lei Liu, Liang-Xiao Ma, Yi-Xue Li, and Yuan-Yuan Li. "DCGL v2. 0: an R package for unveiling differential regulation from differential co-expression." In: *PloS one* 8.11 (2013), e79729.

[594] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, et al. "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells." In: *Nucleic acids research* 41.D1 (2012), pp. D955–D961.

[595] Andrew D Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, et al. "Ensembl 2020." In: *Nucleic acids research* 48.D1 (2020), pp. D682–D688.

[596] Hon Yan Kelvin Yip and Antonella Papa. "Signaling Pathways in Cancer: Therapeutic Targets, Combinatorial Treatments, and New Developments." In: *Cells* 10.3 (2021), p. 659.

[597] TL Yuan and LC Cantley. "PI3K pathway alterations in cancer: variations on a theme." In: *Oncogene* 27.41 (2008), pp. 5497–5510.

[598] Koichi Yuki, Miho Fujiogi, and Sophia Koutsogiannaki. "COVID-19 pathophysiology: A review." In: *Clinical immunology* 215 (2020), p. 108427.

[599] Tahera Zabuawala, David A Taffany, Sudarshana M Sharma, Anand Merchant, Brett Adair, Ruchika Srinivasan, Thomas J Rosol, Soledad Fernandez, Kun Huang, Gustavo Leone, et al. "An ets2-driven transcriptional program in tumor-associated macrophages promotes tumor metastasis." In: *Cancer research* 70.4 (2010), pp. 1323–1333.

[600] Konstantinos Zagganas, Thanasis Vergoulis, Maria D Paraskevopoulou, Ioannis S Vlachos, Spiros Skiadopoulos, and Theodore Dalamagas. "BUFET: boosting the unbiased miRNA functional enrichment analysis using bitsets." In: *BMC bioinformatics* 18.1 (2017), pp. 1–8.

[601] Wojciech Zakrzewski, Maciej Dobrzyński, Maria Szymonowicz, and Zbigniew Rybak. "Stem cells: past, present, and future." In: *Stem cell research & therapy* 10.1 (2019), pp. 1–22.

[602] Ivan Zanoni and Francesca Granucci. "Regulation and dysregulation of innate immunity by NFAT signaling downstream of pattern recognition receptors (PRRs)." In: *European journal of immunology* 42.8 (2012), pp. 1924–1931.

[603] Jerrold H Zar. *Biostatistical analysis*. Pearson Education India, 1999.

[604] Jerrold H Zar. *Biostatistical analysis*. Pearson Education India, 1999. Chap. 3, X–Y.

[605] T Zhan, N Rindtorff, and Michael Boutros. "Wnt signaling in cancer." In: *Oncogene* 36.11 (2017), pp. 1461–1473.

[606] Junjun Zhang, Rosita Bajari, Dusan Andric, Francois Gerthoffert, Alexandru Lepsa, Hardeep Nahal-Bose, Lincoln D Stein, and Vincent Ferretti. "The international cancer genome consortium data portal." In: *Nature biotechnology* 37.4 (2019), pp. 367–369.

[607] Meishan Zhang, Chunming Xu, Diter von Wettstein, and Bao Liu. "Tissue-specific differences in cytosine methylation and their association with differential gene expression in sorghum." In: *Plant physiology* 156.4 (2011), pp. 1955–1966.

[608] Shu-Tao Zhang, Rui Zhao, Wen-Xiang Ma, Yan-Yan Fan, Wen-Zheng Guan, Jiao Wang, Peng Ren, Kun Zhong, Tian-Shui Yu, Jing-Bo Pi, et al. "Nrf1 is time-dependently expressed and distributed in the distinct cell types after trauma to skeletal muscles in rats." In: (2013).

[609] Y Zhang, T Liu, CA Meyer, J Eeckhoute, DS Johnson, BE Bernstein, C Nusbaum, RM Myers, M Brown, W Li, et al. "et al.(2008) Model–based analysis of ChIP–Seq(MACS)." In: *Genome Biol* 9.9 (), R137.

[610] Yan Zhang, Jie Lv, Hongbo Liu, Jiang Zhu, Jianzhong Su, Qiong Wu, Yunfeng Qi, Fang Wang, and Xia Li. "HHMD: the human histone modification database." In: *Nucleic acids research* 38.suppl_1 (2010), pp. D149–D154.

[611] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. "Model-based analysis of ChIP-Seq (MACS)." In: *Genome biology* 9.9 (2008), pp. 1–9.

[612] Jian Zhou and Olga G Troyanskaya. "Predicting effects of noncoding variants with deep learning–based sequence model." In: *Nature methods* 12.10 (2015), pp. 931–934.

[613] Nian Zhou, Bing Yan, Jing Ma, Hongchao Jiang, Li Li, Haoyu Tang, Fengming Ji, and Zhigang Yao. "Expression of TCF3 in Wilms' tumor and its regulatory role in kidney tumor cell viability, migration and apoptosis in vitro." In: *Molecular Medicine Reports* 24.3 (2021), pp. 1–8.

[614] Yonggang Zhou, Binqing Fu, Xiaohu Zheng, Dongsheng Wang, Changcheng Zhao, Yingjie Qi, Rui Sun, Zhigang Tian, Xiaoling Xu, and Haiming Wei. "Pathogenic T-cells and inflammatory monocytes incite inflammatory storms in severe

COVID-19 patients." In: *National Science Review* 7.6 (2020), pp. 998–1002.

[615] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. "Comparative analysis of single-cell RNA sequencing methods." In: *Molecular cell* 65.4 (2017), pp. 631–643.

[616] Quan Zou, Yaozong Mao, Lingling Hu, Yunfeng Wu, and Zhiliang Ji. "miRClassify: an advanced web server for miRNA family classification and annotation." In: *Computers in biology and medicine* 45 (2014), pp. 157–160.

[617] Stanley Zucker, Molin Wang, Joseph A Sparano, William J Gradishar, James N Ingle, and Nancy E Davidson. "Plasma matrix metalloproteinases 7 and 9 in patients with metastatic breast cancer treated with marimastat or placebo: Eastern Cooperative Oncology Group trial E2196." In: *Clinical breast cancer* 6.6 (2006), pp. 525–529.