Integration of public high-throughput databases in quantitative biology

DISSERTATION

zur Erlangung des akademischen Grades
Doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt dem Rat
der Technischen Fakultät
der Albert-Ludwigs-Universität Freiburg

Februar 2020

von
SEBASTIAN OHSE
M.Sc. Bioinformatik und Systembiologie

Dekan:

Prof. Dr. Rolf Backofen

Gutachter:

Prof. Dr. Rolf Backofen

Prof. Dr. Hauke Busch

Datum der Promotion:

26.11.2020

Zusammenfassung

Im gegenwärtigen Zeitalter der biologischen Forschung wird eine beispiellose Menge an quantitativen hochdimensionalen Daten erhoben. Insbesondere im Bereich der Molekular- und Zellbiologie wurden umfangreiche öffentliche Datenbanken eingerichtet, um die Fülle von Hochdurchsatzdaten verfügbar zu machen. Die Integration solcher Daten ist jedoch eine Herausforderung. Um Meta-Analysen durchzuführen oder öffentliche Datenbanken zur Unterstützung einzelner Experimente zu verwenden, ist eine angemessene Normalisierung entscheidend. Auch die Bewertung von Hypothesentests hinsichtlich der biologischen Relevanz ist für hochdimensionale Daten schwierig durchzuführen und erfordert im Allgemeinen Domänenexpertise. Diese beiden Engpässe werden in dieser Arbeit durch die Entwicklung von zwei Algorithmen adressiert: ein Normalisierungsalgorithmus, der systematische Fehlerkorrektur durch die Nutzung von erkennbaren Redundanzen in öffentlichen Datenbanken durchführt und ein empirisches Maß der biologischen Relevanz, welches geeignete Nullverteilungen für verschiedene Teststatistiken bereitstellt. Beide Engpässe wurden durch die Einrichtung von Workflows für die Verarbeitung von Hochdurchsatzdaten im Rahmen von zwei großen Forschungskonsortien identifiziert, die sich mit den Auswirkungen verschiedener Perturbationen auf die Systemeigenschaften von Krankheiten befassen. Insgesamt stellt diese Arbeit einen neuen Blickwinkel auf wichtige Herausforderungen in der quantitativen Biologie da und stellt zwei Algorithmen vor, die in Softwarepaketen implementiert sind um diese anzugehen[1].

[1] http://github.com/a378ec99

Abstract

In the current age of biological research an unprecedented amount of quantitative high-dimensional data is being obtained. Especially in the domain of molecular and cell biology large public databases have been established to make the wealth of high-throughput data acquired available. However, integration of such high-throughput data is challenging. In order to conduct meta-analyses or use public databases in support of individual experiments, appropriate normalization is critical. Furthermore, the evaluation of hypothesis tests with respect to biological relevance is difficult to perform for high-throughput data and generally requires domain expertise. These two bottlenecks are addressed in this thesis through the development of two specific algorithms: a blind normalization algorithm that performs bias correction by leveraging detectable redundancies in public databases and an empirical measure of biological relevance that provides appropriate null distributions for common test statistics. Both bottlenecks were identified through the establishment of workflows for the processing of high-throughput data in the framework of two interdisciplinary research consortia which are concerned with the effect of different perturbations on the systems properties of disease. Overall, this work provides a new view on important challenges in quantitative biology and presents two algorithms that are implemented in software packages[2] to address these.

[2] http://github.com/a378ec99

Publications

A.1 **Ohse**, S., Boerries, M., Busch, H. (2019). Blind normalization of public high-throughput databases. *PeerJ Computer Science*, 5, 231-247.

A.2 Valverde, S., **Ohse**, S., Turalska, M., West, B. J., & Garcia-Ojalvo, J. (2015). Structural determinants of criticality in biological networks. *Frontiers in physiology*, 6(127), 1-9.

B.1 Kalfalah, F., Seggewiß, S., Walter, R., Tigges, J., Moreno-Villanueva, M., Bürkle, A., **Ohse**, S., Busch, H., Boerries, M., Royer-Pokora, B. & Boege, F.(2015). Structural chromosome abnormalities, increased DNA strand breaks and DNA strand break repair deficiency in dermal fibroblasts from old female human donors. *Aging*, 7(2), 110-122.

B.2 Kalfalah, F., Sobek, S., Bornholz, B., Götz-Rösch, C., Tigges, J., Fritsche, E., Krutmann, J., Köhrer, K., Deenen, R., **Ohse**, S., Boerries, M., Busch, H. & Boege, F. (2014). Inadequate mito-biogenesis in primary dermal fibroblasts from old humans is associated with impairment of PGC1A-independent stimulation. *Experimental gerontology*, 56, 59-68.

C.1 **Ohse***, S., Dvornikov*, D., Schneider*, M. A., Szczygieł, M., Titkova, I., Rosenblatt, M., Muley, T., Warth, A., Herth, F. J., Dienemann, H., Thomas, M., Timmer, J., Schilling, M., Busch, H., Boerries, M., Meister, M., & Klingmüller, U. (2018). Expression ratio of the TGFβ-inducible gene MYO10 is prognostic for overall survival of squamous cell lung cancer patients and predicts chemotherapy response. *Scientific Reports*, 8(1), 9517-9530.

C.2 Gladilin, E., Dvornikov, D., **Ohse**, S., Merkle, R., Depner, S., Boerries, M., Busch, H., Meister, L., Schneider, L., Meister, M., Klingmüller, U. & Eils, R. (2018). TGFβ-induced cytoskeletal remodeling mediates elevation of cell stiffness and invasiveness in NSCLCs. *Scientific Reports*, 9(1), 7667-7679.

* These authors contributed equally

Acknowledgements

I am grateful for the support and guidance of my supervisors Prof. Hauke Busch, Prof. Melanie Börries, Prof. Jordi Garcia-Ojalvo and Prof. Rolf Backofen. In addition, I would like to thank the members of the examination committee for their time and effort in evaluating this work. Throughout my PhD I have had uncountable discussions and interactions with many bright scientists from the Center for Biological Systems Analysis, Institute of Molecular Medicine and Cell Research, the department of Computer Science at the University of Freiburg and at the Barcelona Biomedical Research Park. I would like to extend my sincere gratitude to the following friends and colleges for their support, friendship and many stimulating discussions. The following list is incomplete and in no particular order: Aaron Klein, Bence Mélykúti, Fabrizio Costa, Martin Mann, Björn Grünig, Theresa Schredelseker, Teresa Müller, Bérénice Batut, Geoffroy Andrieux, Jochen Hochrein, Hagen Klett, Silke Kowar, Patrick Metzger, Ella Levit-Zerdoun, Lars Nilse, Johannes Hummel, Colin Seibel, Enaam Chleilat, Jie Bao, Sebastian Weber, Moritz Buck, Marisa Fernández-Cachón, Martin Klose, Juliana Nascimento, Andreas Schüttler, Steffen Lemke, Aikaterini Symeonidi, Marçal Gabaldà, Letícia Galera, Alessandro Barardi, Bernat Bramon, Marit Hoffmeyer, Maja Temerinac-Ott, and Matthias Heizmann. To those I missed, you know who you are. My heartfelt thanks goes out to my family who made this journey possible. Last but not least, I would like to thank myself for the sustained dedication and effort that this work represents and the courage to follow through.

"Education is that which remains when we have forgotten all that we have been taught."

Anonymous

Contents

List of Figures

List of Tables

List of Abbreviations

EBI	**E**uropean **B**ioinformatics **I**nstitute
NCBI	**N**ational **C**enter for **B**iotechnology **I**nformation
GEO	**G**ene **E**xpression **O**mnibus
GO	**G**ene **O**nthology
NGS	**N**ext **G**eneration **S**equencing
EST	**E**xpressed **S**equence **T**ag
NLP	**N**atural **L**anguage **P**rocessing
NP-HARD	**N**on-deterministic **P**olynomial-time **hard**
BP	**B**asis **P**ursuit
LASSO	**L**east **A**bsolute **S**hrinkage and **S**election
LARS	**L**east **A**ngle **R**egression
ALS	**A**lternating **L**east **S**quares
IHT	**I**terative **H**ard **T**hresholding
NIHT	**N**ormalized **I**terative **H**ard **T**hresholding
CGIHT	**C**onjugate **G**radient **I**terative **H**ard **T**hresholding
MP	**M**atching **P**ursuit
OMG	**O**rthogonal **M**atching **P**ursuits
CoSaMP	**Co**mpressive **Sa**mpling **M**atching **P**ursuits
ADMiRA	**A**tomic **D**ecomposition for **Mi**nimum **R**ank **A**pproximation
GC-MS	**G**as **C**hromatography based **M**ass **S**pectrometry
LC-MS	**L**iquid **C**hromatography based **M**ass **S**pectrometry
ESI	Electro**S**pray **I**onization
MALDI	**M**atrix-**A**ssisted **L**aser **D**esorbtion or **I**onization
ICAT	**I**sotope **C**oded **A**ffinity **T**ag
iTRAQ	**i**sobaric **T**ag for **R**elative and **A**bsolute **Q**uantification
ICPL	**I**sotope-**C**oded **P**rotein **L**abel
SILAC	**S**table **I**sotope **L**abeling with **A**mino acids in **C**ell culture
TOF	**T**ime **o**f **F**light
AMT	**A**ccurate **M**ass **T**ime tagging
DNA	**D**eoxyribo**n**ucleic **A**cid
RNA	**R**ibo**n**ucleic **A**cid
mRNA	**m**essenger **R**ibo**n**ucleic **A**cid
ncRNA	**n**on-**c**oding **R**ibo**n**ucleic **A**cid
miRNA	**mi**cro **R**ibo**n**ucleic **A**cid
tRNA	**t**ransfer **R**ibo**n**ucleic **A**cid

rRNA	**r**ibosomal **R**ibo**n**ucleic **A**cid
HGF	**H**epatocyte **G**rowth **F**actor
TGFβ	**T**ransforming **G**rowth **F**actor β
IGF	**I**nsulin-like **G**rowth **F**actor
EGF	**E**pidermal **G**rowth **F**actor
EMT	**E**pithelial-**M**esenchymal **T**ransition
BCN	**B**lind **C**ompressive **N**ormalization
MBR	**M**easure of **B**iological **R**elevance
HPC	**H**igh **P**erformance **C**omputing

List of Symbols

$\mathcal{A}$	Measurement operator
$\mathbf{A}$	Measurement matrix
$\mathbf{\Phi}$	Canonical basis
$\mathbf{\Psi}$	Sensing basis
θ	Sparse signal
ρ	Strength of dependence
μ	Strength of coherence
$\mathbf{\Omega}$	Directed graph
H	Hamming distance

Chapter 1

Background

1.1 Quantitative biology

In the current age of biological research an unprecedented amount of quantitative and high-dimensional data is being obtained. Specifically in the domain of molecular and cell biology, as well as medicine, the multitude of molecules which regulate and form the structure of biological systems are now being measured at scale (Joyce and Palsson, 2006). This process appears irreversible and the question thus addressed in this thesis is how the resulting deluge of data can be analyzed and integrated appropriately.

With the rise of high-throughput measurement technologies that allow for the large scale measurement of biological molecules in a quantitative fashion, the traditionally qualitative domains of cell and molecular biology are being transformed: from a phenotype based descriptive science to a quantitative science. Throughout this transition various challenges remain to be surmounted, most importantly the conversion of the massive quantities of data that are being obtained, into scientific knowledge. In Section 1.1.1, a holistic approach to tackle this challenge is outlined for the specific case of the molecular cell. Further, it is discussed how cells can be categorized into different global states, such as homeostasis and differentiation, which in turn are controlled by regulatory networks that determine cell state decisions.

Various novel measurement technologies have been developed, which form the core of what has been termed high-throughput technologies. These include transcriptomics, proteomics and metabolomics and are performed routinely in medical, molecular and cell biological research. Transcriptomics currently measures up to 100,000 genes in parallel when applied to mammalian cells. Proteomics has yet to reach the state where more than 1,000 proteins can be measured consistently and metabolomics is focused on the order of 10-100s of molecules for now. The inner workings of these technologies are described in detail in Section 1.1.2. Current transcriptomics experiments are based on microarray and next-generation sequencing technologies, which measure the global RNA content of cells. Proteomics and

metabolomics experiments, on the other hand, are based on mass spectrometry technology that can identify and quantify the global protein and metabolite content of cells, respectively.

The conversion of massive quantities of data into scientific knowledge requires an appropriate evaluation and integration of the obtained high-throughput data. Thus the development of approaches that map quantitative information into categorical information is crucial. Such data evaluation and integration techniques are limited by two main factors outlined in Section 1.1.3, which discusses current challenges in quantitative biology. The first factor is that the obtained data is generally confounded with bias from various sources, making data integration across experiments and measurement technologies challenging. Secondly, the precise relationship between the measurement scale on which different molecules are measured and their actual biological effect is often only vaguely defined for experiments conducted. Therefore, the interpretation of quantitative comparisons between different molecules is typically limited. Along the same line, null distributions are typically ill defined for the test statistics used to evaluate high-throughput data.

1.1.1 Molecular systems

To bring order into the realm of molecular and cell biology, a holistic approach to the study of molecular systems is needed. This approach must manage the enormous complexity that is the result of a large number of heterogeneous molecules that interact on various levels to produce one of the most fundamental building blocks of life – the cell. Precisely this challenge is tackled by the field of systems biology, which studies both large and small biological systems at scale. By this holistic approach the transformation of biological research to a quantitative science has been very fruitful and has driven the need for more holistic analyses. While systems under studied are not constrained exclusively to molecular systems, the focus is here on the cell as this is where most research activity in quantitative biology is concentrated at the moment.

Methods used in the systematic study of whole molecular systems are typically derived from engineering disciplines and include computational and mathematical modeling approaches. Standard engineering disciplines generally study and categorize parts of complex systems in great detail, to then determine their interactions. An opposing approach is the focus on a holistic understanding of complete systems, such as the cell, tissues, or the human body, that the here discussed systems biological approach strives to understand. Systems biology seeks a combination of the bottom-up approach with the less developed top-down approach to the understanding of complex systems. The goal is to enable the study of potentially synergistic effects that are not apparent by the detailed study of molecular building blocks

themselves, while in addition mapping the structure of the molecular system as a whole.

One milestone reached by systems biology has been the creation of a whole cell model of bacteria (Karr et al., 2012). This whole cell model is able to predict the viability of the organism under different perturbations, such as genetic mutations (Karr et al., 2012). A simulator for gene expression changes over time has also been developed, termed the Gene Net Weaver (Schaffter, Marbach, and Floreano, 2011). A particular focus of systems biology involves the study of networks, such as gene regulatory networks or other types of cell signaling networks. Here systems and emergent properties, such as criticality, have been characterized through the application of dynamical approaches of systems biology, that aim to understand fundamental processes that occur in the cell (Valverde et al., 2015).

Challenges in the area of systems biology include a current downturn in funding and the difficulty in evaluating more complex models, especially when it comes to the analysis of complex qualitative phenotypes. At the moment the characterization of cellular molecules has been more successful than the description of their holistic dynamics. However, with respect to funding agency goals in support of the development of new treatments, this has been disappointing. Systems biology has also started to extract and categorize information from sources such as text and high-throughput data. For this purpose online repositories have been developed to store this data publicly. Furthermore, the process of developing meta level standards for the description and exchange of statistical models and interacting parts of cellular subsystems has begun. For example, the standard registry of biological parts which arose out of the iGEM competition has been leading standardization in the field of systems biology (iGEM, 2017).

What is a complex system? It is easiest to define a complex system as what it is not – simple to predict. Naturally, complex systems have been studied empirically since the beginning of science. But, in the past decades it has become clear that complexity itself arises from the nature of our description of the physical world and is not only due to incomplete information about a difficult phenomenon, or models which contain too many parameters relative to limited empirical evidence. Complex systems generally contain many interacting parts that may lead to emergent phenomena not initially apparent from the parts themselves. Such effects are termed synergistic and arise in models of the weather, large scale social or economic interactions, the electro-physiology of the brain and the stock market; mostly any system that has a certain level of complexity. Standard scientific practices based on reductionism quickly reach their limits when confronted with such systems and new strategies need to be developed.

The cell can be understood as a complex system (Ziemelis, 2001). Hence, reductionist techniques aimed at understanding its internal workings may prove challenging. Such techniques have been successfully applied in the past to biological problems, but there are major stumbling blocks once the complexity of a particular systems biological research a certain level. For example, synergistic effects can arise when many small parts are interconnected to produce effects that can not easily be inferred from the parts themselves. Specifically, the phenomenon of criticality has long been noted as one of these stumbling blocks (Valverde et al., 2015). Today, many questions in molecular and cell biology are concerned with highly specific parts of the cell that nonetheless have a large effect on the whole cell. For example, a major question is which type of molecular stimulus does give a desired cellular change? Models of the whole cell are clearly not yet able to predict such dynamics changes. Only recently has it been possible to measure many of the components that underly these phenomena and construct models based on them (Reuter, Spacek, and Snyder, 2015). There remains much to be understood from a complex systems perspective with respect to the dynamics of the cell. Currently, cellular states can be categorized into two main groups, homeostasis and differentiation, described in the next section.

Cellular states

Homeostasis

Homeostasis is the process of maintaining a stable steady state. The cell requires constant regulation of its internal processes and external cell functions in order to sustain such a steady state for itself and the encompassing organism. A major component of homeostasis is the regulation of metabolic processes, which impacts all cellular functions. The ability to adjust anabolic and catabolic processes is critical for maintaining an appropriate energy balance. Such control is mediated through active and passive processes, including protein phosphorylation, binding of catalysts or gene expression changes. In both single and multi-cellular systems the appropriate regulation of metabolism reflects the abundance of nutrients relative to the internal energy balance . Thus, cells necessarily have evolved an ability to sense their external microenvironment through the use of signaling systems, such as receptor based detection specific signaling molecules and subsequent downstream activation of the corresponding effectors. These signal transduction processes commonly make use of phosphorylation or other modifications of macromolecules to pass on a specific signal. This can result in signaling cascades as shown in Figure 1.1.

Differentiation

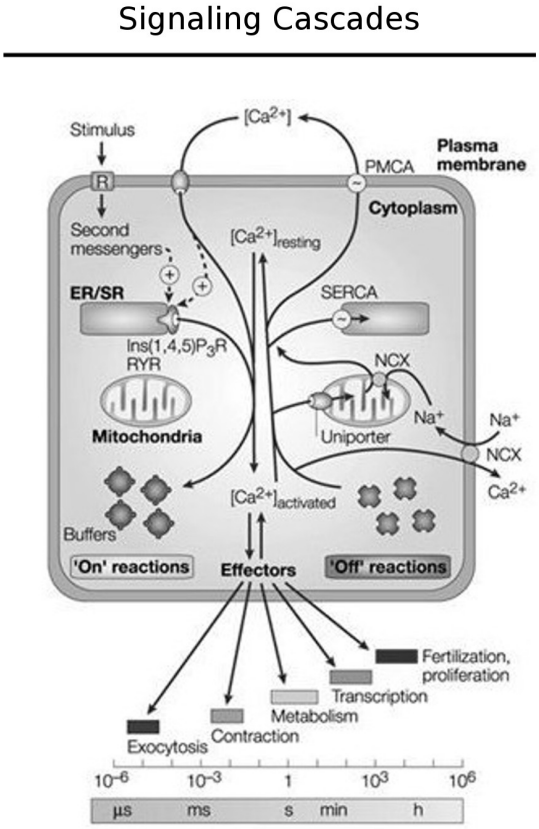


FIGURE 1.1: Example of a signaling cascades in the mammalian cell. Calcium signaling induces downstream effectors. Modified with permission (Berridge, Bootman, and Roderick, 2003).

The cell and its ability to differentiate into different cell types is the epitome of a complex system and the main focus of systems biology. Typically, the process of differentiation starts from a precursor or more general cell type to a more specific cell subtype during normal developmental or regenerative processes. An early and widely studied model is that of Waddington's landscape of cell differentiation (see Figure 1.2. However, differentiation processes are not exclusively in a direction of more specificity or more constrained cellular plasticity. The cellular characteristics that change during differentiation include modification of the cellular membrane, the molecular or chemical species produced in the cellular metabolism and deeper changes on the chromatin level. At no point is the DNA sequence information changed other than in rare special cases, such as immune cells. Thus, all cell types in an organism contain an identical DNA sequence. The differences that eventually lead to distinct cell types are only based on gene expression or related epigenetic changes, which subsequently lead to the modifications observed on the structural level. The differentiation changes induced are typically long term and distinct from metabolic or signaling processes on the protein level, such as phosphorylation and methylation, that are targeted more at governing the the regulation of proteins and

enzymes.

Stem cells that can differentiate into all subtypes of cells in an organism are termed pluripotent. Or, in the case of zygote blastomeres, these are termed totipotent. Pluripotent stem cells can be induced from already differentiated cells though a cocktail of four transcription factors termed the Yamanaka factors (Oct4, Sox2, c-Myc, Kfl4) (Takahashi and Yamanaka, 2006). A subset of stem cells that can only differentiate into cell types very similar to the parent cell are termed multipotent stem cells. These are typically found in tissues or organs, such as the skin or the blood stream, which require constant regeneration due to contact with the outside environment or limited lifetime of cells due to other factors. It is hypothesized that in tumors there also exist stem cells, which are the source of new tumor formation after common therapies that remove the bulk of the tumor (Singh et al., 2003). A common measure of tumor grade is how differentiated the tumor is and thus how distinct from a stem cell state. Overall, cell differentiation is an important process that is yet poorly understood on the systems level.

Only recently have efforts been successful at reprogramming cell types in vivo for therapeutic purposes (Naldini, 2015). This process is likely to improve with the establishment of new gene editing tools, such as the CRISPR technology and other advances in the area of gene therapy. In oder to understand how to appropriately reprogram cells, it is important to measure all the genes, proteins and molecules involved, to get a holistic view of what occurs on the systems level. For such measurements high-throughput technologies are needed, which are described in detail in the next sections.

Regulatory networks

In the following section the regulatory networks which govern changes in cellular states are discussed. An example regulatory network is shown in Figure 1.3. A particular focus is placed on the peculiarities of gene regulatory networks, as changes in gene expression underly most of the long-term changes that lead to differentiated cell states. First, the framework of random Boolean networks is introduced and empirical and analytical measures of criticality are discussed. Subsequently, the known relations between critical dynamics and network structure are highlighted and respective evolutionary origins are explored. Most of the material in this section is taken verbatim or in modified form from publication A.2 (Valverde et al., 2015).

Random Boolean Networks

Gene regulatory networks operate on a scale of the order of 10^4 nodes (Hecker, 2009). In order to explore the dynamics within such a large network, a majority of studies

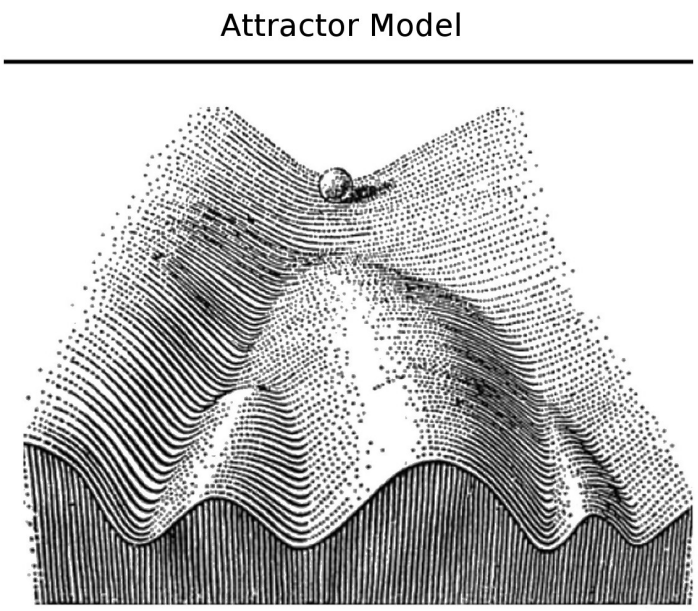


FIGURE 1.2: Waddington's landscape of cell differentiation. The marble on top represents a cell state. Different pathways are possible, but after a choice is made, the subsequent cell states are separated by a barrier. Differentiation is unidirectional according to this model. Modified with permission (Goldberg, Allis, and Bernstein, 2007).

on gene regulatory network dynamics have been conducted within the framework of random Boolean networks (RBNs). This framework was introduced in the late 1960s by Stuart Kauffman, with the specific aim to study the properties of gene regulatory networks (Kauffman, 1969). See Drossel (2008) for a comprehensive review. Briefly, random Boolean networks are a type of complex network with a limited set of allowed node states and transfer functions[1]. The state of each node (gene) is restricted to only two possibilities, on or off. Formally, a Boolean network is a directed graph $\Omega(V, E, B)$ with a set of Boolean functions $B = \{b_i | i = 1 \dots n\}$ such that $b_i : \{0, 1\}^k \rightarrow \{0, 1\}$, with $k \leq n$. Before a simulation in this framework is initiated, a random initial state is set for each node. During the simulation, the state of a node at time t is given by $x_i(t)$, and the next state after each iteration is given by $x_i(t + 1) = b_i(x_{i1}(t), x_{i2}(t) \dots x_{ik}(t))$, where x_{ij} are the states of the nodes connected to node i. The states of all nodes are updated simultaneously according to this rule. This process may be iterated until convergence to a stable fixed point or limit-cycle. The simplification provided by a random Boolean network model enables a systematic exploration of the relationship between network structure and critical dynamics that might otherwise be unfeasible.

[1]Transfer functions integrate the signal from all incoming edges to a particular node to determine the node state at time t.

Measures of criticality

In its most simple realization, each node of a Boolean network is connected at random to a set of K input node, and one chooses uniformly at random a possible transfer function. In such a homogeneous network the median cycle length is $0.5 \cdot 2^{N/2}$ (Legenstein and Maass, 2007). Due to the finite size of the network its convergence is guaranteed. The Hamming distance measures the minimum number of substitutions to convert one Boolean network state into another and is used to measure the evolution of the system at each iteration. Formally:

$$H(t) = \sum_{i=1}^{n} |x_i(t) - \tilde{x}_i(t)|$$

where x and $\tilde{x}$ are two slightly different initial states of the same network and i runs over all nodes of the network (Pomerance et al., 2009). As the number of iterations tends to infinity in a finite network, $H(t) \to 0$, but does so more slowly the more erratic the behavior is. In the limit of infinite size the network can become chaotic. The slope of the $H(t)$ curve at the origin is indicative of criticality. This is an empirical measure (Legenstein and Maass, 2007). According to this measure and under the annealed approximation (through which all Boolean functions are randomized at each iteration), the dynamics becomes critical for $K = 2$, whereas networks with $K = 1$ operate in an ordered regime.

The analytical definitions of criticality are increasingly generalized to allow application to more realistic and complex models of gene regulatory networks. Shmulevich, Kauffman, and Aldana (2005) generalized the initial formula to allow computation for the case where network functions are generated according to probability distributions that favor some variables over others, measured through their activities, or when transfer functions are chosen at random from certain classes (such as canalizing functions). Pomerance et al. (2009) generalized the initial formula to allow (i) any network topology, (ii) a distribution of biases instead of one parameter, (iii) non-synchronous updates, and (iv) multiple node states, while still permitting the calculation of the control parameter at which the network dynamic is critical. The method uses the maximum eigenvalue of a modified adjacency matrix. In any case, it must be noted that the concept of criticality loses its utility without a clear definition of how close to the critical threshold network dynamics must be in order to qualify as critical.

Structural determinants

A major aim in the literature has been to demonstrate the phenomenon of criticality in specific gene regulatory networks, which have been inferred from (incomplete) empirical evidence (Shmulevich, Kauffman, and Aldana, 2005; Nykter et al., 2008;

Balleza et al., 2008). The question of how structural features contribute to the emergence of criticality remains largely unaddressed. Here we give an overview on what is known of the effect of structural properties on the location of critical points within the framework of the random Boolean network model, discussing in particular the case of scale-free architectures and the roles of community structure and canalizing functions.

Scale-free topology

Aldana et al. (2007) argue that a scale-free topology diminishes the need to fine-tune connectivity parameters (the rewiring probability p and the average degree K) to obtain critical dynamics. In particular, the critical phase transition in scale-free networks occurs over a range of scale-free exponents ($\alpha \in [2.0, 2.5]$) and allows for a range of connectivities. Along this line, Fox and Hill (2001) argue that homogeneous topologies with biologically realistic connectivities would lie in the chaotic regime, since their average connectivity (measured by K) is relatively high. If gene regulatory networks indeed operate at criticality, a scale-free topology might explain this discrepancy. In the thermodynamical limit, broad degree distributions do not affect the critical point (provided K is fixed), but in finite settings power-law distributions lead to increased order. For example, even if the average K is large in a given network, there can be many nodes with low in-degrees that are likely to be frozen nodes. This reduces the size of the network that is active and effectively involved in the dynamics, which in turn reduces the real value of average K for the network, since many of those in-degree links might come from frozen connections, and thus do not contribute to potentially chaotic dynamics (Fox and Hill, 2001).

Modularity

The presence of community structure in the network impedes signal transmission, pushing the system into an ordered phase (Wang and Albert, 2013). Also, modularity broadens the range of connectivities which allows for critical dynamics. Modular RBNs have more attractors and are closer to criticality when chaotic dynamics would be expected, compared to classical RBNs (Poblanno-Balp and Gershenson, 2011). In general, modules make it difficult for damage to spread through the network, even if the local connectivity (within a module) is high. In this way, chaotic dynamics can be constrained within modules (Gershenson, 2012). In contrast with the effects described above, modularity also allows for information flow between modules, and thus while reducing the occurrence of chaos it might also contribute to the spreading of the critical regime, much like Griffiths phases (Hesse and Gross, 2014) (see above), because modules are often connected with each other, leading to a small-world topology which in turn allows for more critical dynamics. Lizier, Pritam, and Prokopenko (2011) argue that a small-world topology in RBNs has relatively large

information storage and transfer capabilities and enables critical dynamics.

Transfer functions

The effect of varying the rewiring probability p or their in-degree K can be replicated by changing the incidence of canalizing functions (Shmulevich and Kauffman, 2004), which yield dominating inputs in transfer functions (so that the node would be unaffected by other inputs). Canalizing functions are found with high probability when selecting Boolean functions uniformly at random (Serra, Villani, and Semeria, 2004), and are thought to occur in realistic gene regulatory networks (Shmulevich and Kauffman, 2004). Balleza et al. (2008) used networks from several model organism networks to argue that increasing the probability of canalizing functions, while generally pushing dynamics towards the ordered phase, is not sufficient to leave the critical regime. The results are essentially the same if the fraction of canalizing functions is not inferred from the microarray data (Balleza et al., 2008). The effect may be similar to silencing, the fixation of a subset of nodes in a particular state, which has been shown to make the system more ordered (Serra, Villani, and Semeria, 2004; Luque and Solé, 1997).

Evolutionary mechanisms

Several works have investigated the evolutionary mechanisms leading to network structures that may in turn facilitate critical dynamics (Bornholdt and Rohlf, 2000; Solé and Valverde, 2006; Aldana et al., 2007; Solé and Valverde, 2008; Torres-Sosa, Huang, and Aldana, 2012). Criticality in gene regulatory networks may in fact be ubiquitous due to evolutionary mechanisms. Biological networks are subject to an evolutionary trade-off between conserving essential network function while allowing for modifications that may increase fitness. Clearly, any system replicating and competing under natural selection must be able to conserve current functions; but also needs to be able to adapt. Given these two constraints, Torres-Sosa, Huang, and Aldana (2012) simulate the evolution of gene regulatory networks in the random Boolean framework described above, under a fitness function that penalizes the loss of existing attractors and rewards the creation of novel attractors. Specifically, gene regulatory interactions are mutated and grown by the mechanism of gene duplication. Network instances are selected to maintain their current dynamical attractors (e.g. their current phenotypes) while generating new ones. The authors show that the selected networks display criticality. However, it should be noted that to produce non-trivial networks it is necessary to introduce an α-fitness criterion, which prescribes a low fitness to nodes that are always frozen and thus have a minimal dynamic range.

Another example showing how standard evolutionary mechanisms lead to critical dynamics was given by Bornholdt and Rohlf (2000). The selection rules used in that case were such that nodes that do not change their state within the attractor trajectory receive new connections at every iteration. This leads to an average connectivity of the network equal to the critical connectivity, without the need of tuning the system. In this way this process leads to self-organization of the network in terms of its average connectivity. A similar conclusion was reached by Aldana et al., 2007.

One way to deal with the challenges that criticality represents to the study of regulatory networks is to simply obtain more detailed measurements of the molecular system under study. However, due to the fact that small initial changes in a non-linear system can lead to large changes further down, accurate predictions of cell states are likely to continue to be difficult. The recent development of high-throughput measurement technologies is thus only a step in the right direction but not sufficient.

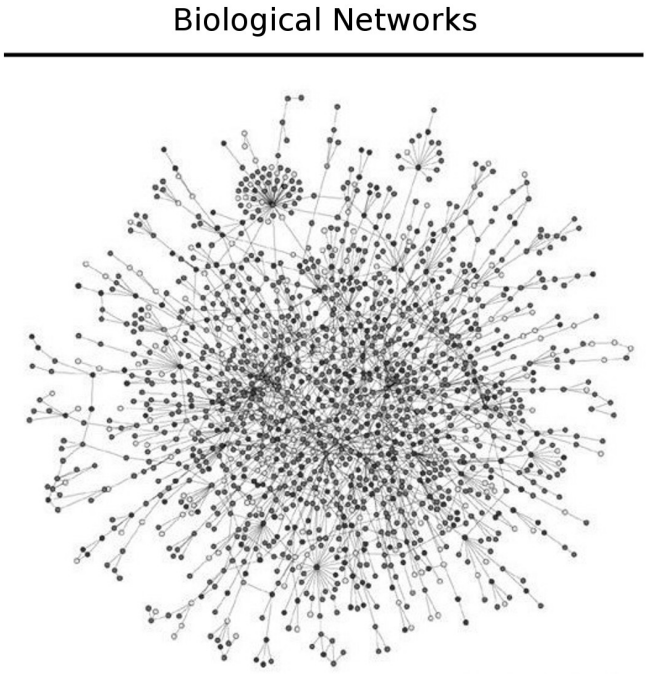


FIGURE 1.3: An undirected protein regulatory network of an eukaryotic cell. Edges represent interactions and nodes denote proteins. High degree nodes are those with most edges and are thought to be involved in critical cell decisions. Modified with permission (Barabasi and Oltvai, 2004).

1.1.2 Measurement technologies

The central dogma of biology states that information in the cell flows from DNA to RNA to protein. See Figure 1.4. It was established in the 1970s through the work of Francis Crick (Crick, 1970). Only recently have measurement technologies caught up to be able to detect these important molecular species at scale. High-throughput technologies can now detect and to some extend quantify a large fraction of the DNA, RNA and protein macromolecules, which exist within the lipid compartment of the cell. In most cases, measurement techniques are not applied on single cells, but on a mixture of millions of cells, due to constraints in the isolation process, which then yield only average measurements.

Measurement technologies specifically focused on the molecular species considered by the central dogma of biology are: genomics (DNA), transcriptomics (RNA), and proteomics (protein). However, there already exist many different combinations, modifications and extensions of these technologies. Cutting-edge examples include 3C technologies, such as ChIA-PET and Hi-C, which are able to capture chromatin conformations that allow a spatial reconstruction of DNA locations within the cell.

Often advances in science closely follow technological innovation. The fact that one can now measure a large portion of the building blocks and regulatory elements of the cell is unprecedented and surely such an advance. Much of the acquired high-throughput data is submitted to public databases and openly accessible (see Figure 1.5). Mainly this is due to the requirements of academic journals and funding agencies. The availability of the produced data from these new technologies to a wide number of scientists is another important factor likely to contribute to major advances in the field of molecular and cell biology. In this section gives a more detailed description of the measurement technologies directly involved in quantifying the molecules involved in the central dogma of biology.

Transcriptomics

Transcriptomics involves the measurement of RNA macromolecules that are produced by the process of transcription, in which an RNA polymerase transcribes DNA into RNA. The major components of cellular RNA species are messenger RNA (mRNA) and non-coding RNA (ncRNA), the later which includes micro RNA (miRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). These macromolecules can give fundamental insights about long term cellular states and dynamics that are occurring within a cell. For example, during cellular differentiation transcription factors bind to different sections of the chromatin (coiled DNA) in order to activate or deactivate the transcription of specific mRNA molecules, which are necessary for the

[2]Wikimedia-Commons (2008), *Central dogma of biology*. Sourced on 26/06/17.

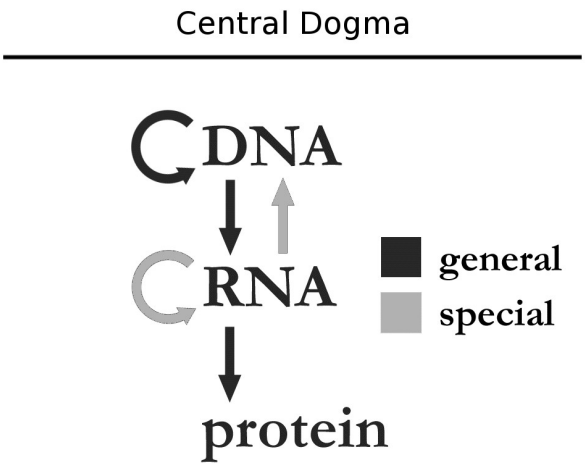


FIGURE 1.4: The central dogma of molecular biology. Information flows from DNA to RNA to protein. However, cross-interactions between DNA, RNA and RNA to DNA have been observed after establishment of the dogma. Modified with permissions [2].

dynamics and functioning of a particular cell type. More details about cell types can be found in the preceding section.

Current transcriptomics measurements are performed with array technologies (e.g. microarray), next-generation sequencing technology (e.g. RNAseq) and expressed sequence tags technology (e.g. EST profiling). Furthermore, within the different measurement technologies there exist different instruments and/or platforms that have been developed. These have various differences such that measurements from one platform can not always be directly compared with another. For example, this may be due to different numbers of transcripts that are measured. Different platforms produced and marketed by different companies for RNASeq include: 454 Life Sciences, llumina, SOLiD, Ion Torrent and PacBio. Illumina has by far the largest market share with around half a million experiments already deposited in the public domain (Barrett et al., 2013). For microarray based technologies the arrays were initially constructed individually for specific experiments. However, reproducibility increased once Affymetrix and Illumina developed more standardized frameworks such as Affymetrix GeneChips and Illumina BeadArrays (Barnes et al., 2005) produced en masse. While methods based on parallel qPCR have been developed, these have been surpassed by next-generation sequencing technology (NGS), which is currently dominating the field (see Figure 1.5).

Proteomics

In proteomics, the protein content of cells is characterized by the identification and quantification of peptides that make up cellular proteins. Typically, the measured

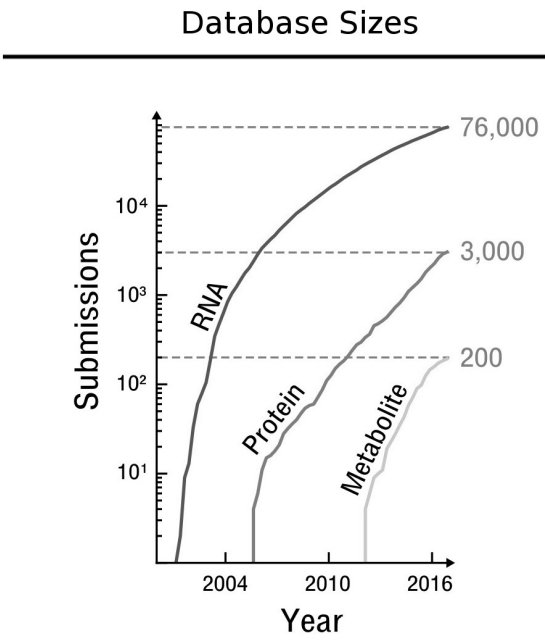


FIGURE 1.5: Rise of high-throughput measurement technologies . RNA submissions are based on NCBI's Gene Expression Omnibus (Barrett et al., 2013), protein on EBI's PRIDE database (Vizcaíno et al., 2016) and metabolite on EBI's MetaboLights database (Haug et al., 2012). Actual samples are approximately an order of magnitude larger than the number of recorded submissions, which typically contain 10-100 samples per submission. The more recent technologies are displaying a rapid growth.

peptides are a result of a chemical or physical degradation of proteins that happens during the initial part of the measurement process (in the bottom-up approach to proteomics). This step is then followed by chromatographic fractionation and subsequent mass/charge determination in a mass spectrometer. Cutting-edge technologies include fractionation based on gas or liquid phase chromatography and the subsequent fragmentation based on a quadrupole. The final measurements of the preprocessed and isolated peptides are conducted in a time of flight mass analyzer (TOF), orbitrap mass spectrometer or quadrupole mass analyzer (QMS). These can be used with gas phase sample injection (GC-MS) or liquid phase sample injection (LC-MS). If at least one additional step of fragmentation is performed by a quadrupole in the instrument setup this type of measurement process is referred to as tandem mass spectrometry. Appropriate peptide ionization, which is required for the mass analyzer, can be performed via electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI). There are several common approaches for mass spectrometry to quantify peptides. For example, isotope coded affinity

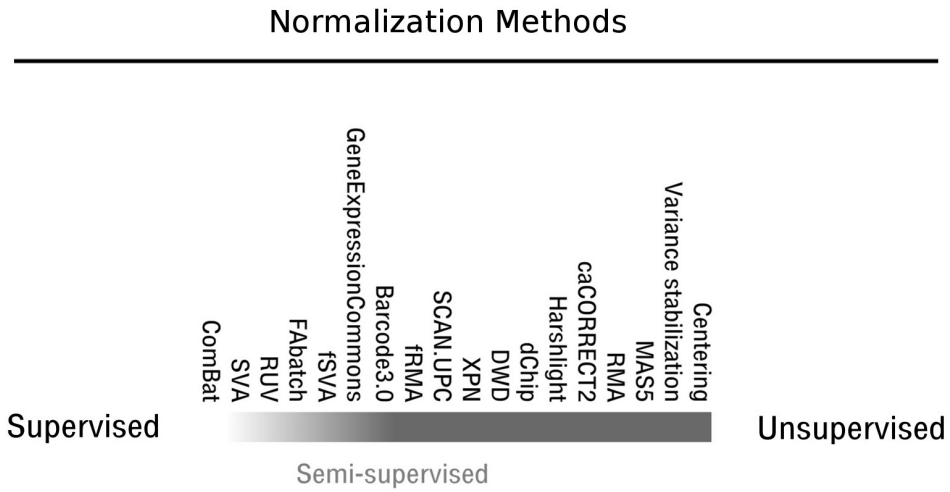


FIGURE 1.6: Overview of current normalization methods from unsupervised to supervised learning (Ohse, Börries, and Busch, 2019).

tag (ICAT), isobaric tag for relative and absolute quantitation (iTRAQ), isotope-coded protein label (ICPL) and stable isotope labeling with amino acids in cell culture (SILAC). Label free quantitative proteomics also exists. These include techniques such as two-dimensional gel electrophoresis (2DE) and accurate mass and time (AMT) tagging.

The detection of petites in proteomics is not as sensitive nor as specific as cutting-edge transcriptome technologies and there are still many less abundant proteins that can not be characterized. Moreover, proteins are subject to various modifications after translation, such as phosphorylation, ubiquitination and several less well studied post-translational modifications, such as methylation, acetylation, glycosylation, oxidation and nitrosylation, which play important roles in signaling cascades (see Figure 1.1). These can also be detected with cutting-edge proteomics technologies, but only when existing in large quantities. There is also a trade-off for proteomics measurement instruments between the time spent on the identification of many different peptides (sensitivity) and the amount of time it spends on the detailed characterization of those peptides (specificity). Thus, depending on the scientific question asked, different strategies need to be pursued. In particular, the decision between a targeted and very sensitive approach that detects and quantifies specific proteins of interest and a more broad exploratory approach that only detects highly concentrated proteins needs to be made.

Metabolomics

Metabolomics involves the collective study of many of the smaller molecules (less than 1 kDa) found in the cell. These make up the cell's metabolic building blocks, including reactants and products of enzymatic processes. Common molecular species characterized in metabolomics workflows include amino acids, glucose and its derivatives, fatty acids, TCA components and on the order of 100 other molecules. More recently, the real time flux of chemical reactions can also be quantified in a high-throughput fashion (Link et al., 2015). Typically, however, experiments are performed with standard GC-MS. Thus, the experimental apparatus is similar as in proteomics. In a standard experiment, a mass spectrometry device is coupled to a liquid or gas chromatography device that fractionates the input. Then, in the typical tandem mass spectrometry setup, further detection and isolation of certain fractions is performed. Molecules that are detectable with these fractions range on the order of 70 Da to 1 kDa.

Of importance is the initial extraction protocol which is typically based on a polar solvent, but can also be non-polar for the case of the isolation of lipid compounds. This makes a large difference for the types of molecules that can be detected by the measurement instrument. The final component of the mass spectrometer is typically an Orbitrap and a quadpol or time of flight (TOF) device. Typically, a mass spectrometry instrument can contain various arrangements of these components, for example 3-4 quadpols followed by a TOF. The obtained data is output as molecular spectra and only semi-quantitative. A common software used to visualize such data is MetaboAnalyst 3.0 (Xia et al., 2012). In addition, much as in proteomics, databases are needed to annotate and identify molecular species. These are METLIN for LC-MS and HMDB, NIST and FIEHN for GC-MS.

Typical normalization routines include an interior standard, which ideally is already put into the medium of the cells to be measured. This can help to get at sample preparation bias and instrument bias and any result is subsequently divided by this factor. The next stage is a blank measurement of the medium or solvent, which is subtracted from the final result. Then, typically a normalization to cell count, or pseudo-cell count based on the sum of peaks in the measured spectra is performed. The matching of compound spectra for identification with those in a databases is based on certain thresholds, such as 5% retention time and other quality characteristics.

Limitations of metabolomics techniques are that currently not all of the complete metabolome can be quantified consistently; only to the degree of pmol to nmol can molecules be measured consistently with cutting-edge setups. The precise sensitivity threshold depends on the type of molecule among other factors. While this is the most nascent and therefore least developed of the three omics technologies

discussed, it allows for the reconstruction of a more complete picture of the central dogma of molecular biology (see Figure 1.4). Not only are RNA and protein levels then measured, but also the reactants and products of enzymatic reactions that underly the processes that contribute to the functioning of the cell and its non-equilibrium state.

1.1.3 Challenges

The challenges found in quantitative biology are common to other areas of science. Particularly, in dealing with high-throughput technologies, a common issue is that of limited funding or statistical expertise. Without the later, it is difficult to perform appropriate experimental controls and replication, which then leads to a low signal-to-noise ratio in the resulting data. This makes subsequent downstream analyses and biological interpretation difficult to perform and less informative once completed. Typically, some *post hoc* fix needs to be applied by those conducting the downstream analysis, which is much more challenging than designing experiments appropriately in the first place. It is difficult to address such concerns, especially when these are exacerbated due to vague or exploratory research questions frequently found in systems biological research. Such shortcomings can not easily be surpassed by advances in method development, but can only be addressed through general scientific education in the area of experimental design and statistics. Here, the peer review process plays an important role, when deciding on the criteria that is used to evaluate which studies have been conducted appropriately. More emphasis in this area should push the community to adhere to higher standards in the initial experimental design.

Tackled in this work are two challenges of quantitative biology that can be addressed by methodological improvements. First, the prevalence of biases due to confounding factors, such as those stemming from differences in measurement instruments (or experimental setup), experimental designs or a lack of normalization are addressed. In particular, the *post hoc* reduction of such shortcomings is critical for the integrating high-throughput experiments deposited in the public domain. Secondly, the current lack of a measure of biological relevance is addressed. For such a measure, the measurement scale used is important when it comes to integrated features describing one phenotype. Without a proper measure of biological relevance, an appropriate prioritization of experimental findings in biological findings is intangible when analyzing high-throughput experiments. Both these methodological challenges are outlined in the next sections and potential solutions to them are formulated in the next chapters. The next sections are take in part from publication A.1 (Ohse, Börries, and Busch, 2019).

Confounding factors

In standard experiments only a particular factor is varied and ideally the remaining variables are kept constant. This approach lies at the heart of the scientific method and is based on a strategy of compartmentalized and controlled experiments. However, this approach becomes challenging once the system under study reaches a complexity, where in the case of high-throughput data, the high-dimensional measurements of molecular species obtained have so many interactions and synergistic effects that compartmentalization becomes unfeasible. Thus, in the area of systems biological research, where complex organisms are studies, it is not always feasible to keep everything relevant under consideration such that the experiment can be said to be controlled. Therefore, biological confounding factors in addition to technical factors are a common occurrence in high-throughput experiments. These systematics biases, sometimes termed batch effects, limit the interpretation of analyses and can be understood as confounding factors (Lazar et al., 2012).

Apart from this fundamental limitation, the measurement process itself also leads to the occurrence of confounding factors. The parallel measurement technology that is used to obtain large scale measurements in a high-throughput fashion has the drawback that the measured molecules can become influenced by each other during the measurement process. For example, some species of molecules may drown out the signal of others and a measurement instrument can typically only be calibrated optimally for a particular subset of molecules of interest. Due to the almost nonexistent use of measurement standards, no simple correction of these effects is possible. Normalization of such effects *post hoc* is challenging and an active field of active research, as described in the next section. In some cases normalization by a standard may not even be possible, since there exists no standard cell to be used as positive control or standard. Even if such standard cells existed, it would require an exorbitant number of combinations of different standard cells to cover the whole spectrum of potential measurements. Thus, confounding factors are difficult to correct in modern high-throughput technologies.

Normalization, or more precisely standardization, involves the attenuation of bias resulting from confounding factors affecting the measurement process. Technical bias of an instrument or sample preparation procedure can be addressed by measuring identically processed standards of known quantity. Use of such standards is widespread in serial technologies. The further up-stream in the measurement process standards are introduced, the more potential sources of bias can be accounted for. Biological bias due to non-identical cells or organisms is often addressed by randomization instead (Montgomery, 2008). This presupposes that the contrast of interest and potential bias sources are already known, as is often the case for individual experiments with performed with serial technologies. An overview of potential bias sources is given by Lazar et al., 2012 and shown in Figure 1.7.

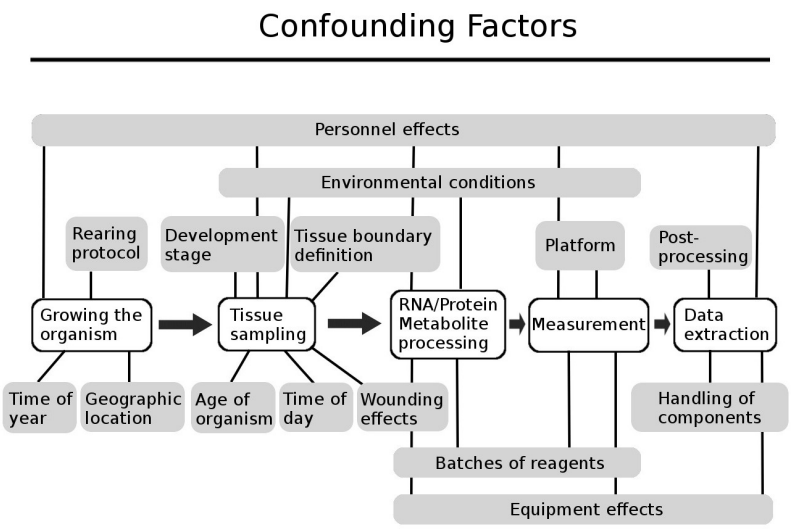


FIGURE 1.7: Overview of confounding factors in high-throughput measurements and databases. The factors considered here are typically termed *batch effects*. Modified with permission from (Lazar et al., 2012).

High-throughput technologies are challenging to standardize in part because the bias of biological molecules measured in parallel is not sample independent, as mentioned in the preceding section. The dependent bias stems from interactions throughout the measurement process, including sample preparation procedures and instrument settings that are dependent on the measured sample and its biological signal. Potential measurement standards must therefore effectively cover a vast number of possible combinations of different quantities. In addition, instrument or measurement process components are sometimes one-time-use, such as in the case of microarray technology, making appropriate prior standardization impossible. In part for these reasons, high-throughput technologies have been initially designed with a focus on relative comparisons, such as fold changes rather than absolute quantification. While a limited number of spike-in standards can account for some technical bias (Lovén et al., 2012), sample preparation procedures that are important sources of bias, such as library preparation, protein extraction, or metabolic labeling, generally happen up-stream of spike-in addition. Bias attenuation by randomization is also not generally possible, as contrasts of interest are not initially known in the exploratory analyses that are typically performed with high-throughput technologies.

The initial experimental design establishes how standards and randomization are employed in a particular experiment. However, in the case of experiments that draw on public databases, the attenuation of bias must be done *post hoc*. Attempts at

such *post hoc* normalization have produced methods across the spectrum of unsupervised to supervised learning shown in Figure 1.6. Unsupervised approaches make use of ad hoc assumptions about noise sources or biological signal, which are then leveraged in an attempt to average out bias. However, unsupervised approaches fail to exploit the wealth of information contained in high-throughput databases and it is difficult to assess the appropriateness of their underlying assumptions for a particular experiment. Supervised approaches make use of prior knowledge of potential confounding factors and contrasts of interest to perform *post hoc* bias attenuation. But, these methods are unfeasible in the case of large databases with insufficient or incoherent annotation and unknown contrasts of interest. Semi-supervised approaches have been introduced that implicitly or explicitly aim to exploit additional high-throughput data to learn parameters that can be transferred. However, current techniques require knowledge of contrasts of interest for the additional data to be of use, or are only concerned with rescaling, or are only possible for the case of normalizing between two known bias sources. Overall, the exploitation of public high-throughput databases for semi-supervised normalization approaches only remains to become more prominent and effective as database sizes continue to increase.

In Chapter 2, an unsupervised approach to the normalization of high-throughput databases is proposed, which does not rely on additional information about sample contrasts or other ad hoc assumptions about the structure of confounding bias. It is thus a blind recovery approach, that does not need one to specify a prior model of the underlying confounders. Instead, the algorithm leverages detectable redundancies in the high-throughput database to be normalized. These redundancies may be similar features or samples. Additional information, such as spike-in standards can be combined into the bias recovery framework and missing values are supported. The assumptions necessary for the approach to be applicable are discussed in Chapter 2.

Measurement scales

In the current transition from qualitative to quantitative biology what often remains behind is a clear definition of the measurement scale. In order to interpret and draw biological meaning from quantitative measurements the definition and meaning of a particular measurement scale used is crucial (see Table 1.1). However, within high-throughput technologies a common measure has been the fold change between different experimental conditions. Only recently have absolute quantitative measurements been introduced. In the area of transcriptomics it is often not clear how to summarize different gene expression measurements into gene sets appropriately, since the appropriate scale for each gene may be quite different and typically not known. It is not clear how to relate the measurement scale of complex systems to the phenotypes observed and thus how to interpret the biological significance of, for

example, a large change. Therefore, it is important to normalize the scales of individual genes to be comparable. But, as discussed, the challenge is that depending on the experimental setup and the biological phenomenon studied, the appropriate measurement scale may be different.

In the past, the definition of an appropriate measurement scale has been answered through the careful weighting of pilot experiments and the clear definition of the particular phenotype of interest. Currently there are no such efforts underway as the precise phenotypical definition has been less of an issue in small-scale experiments, but is challenging for high-throughput experiments, especially with respect to exploratory studies. Also, complex systems such as the cell were not studied with a top-down approach in the past, as most attention was paid to the characterization of individual parts and components. Therefore, this particular challenge is coming to the forefront with the advent of high-throughput technology. Especially when there are incoherent quantitative phenotypes of equal importance within one experiment, a clear definition of the measurement scale is challenging. It may thus be up to each individual study to clearly define its goals with respect to a particular experiment, which can then support the choice of specific quantitative phenotype over another that is then used to determine the appropriate measurement scale. However, as a consequence comparisons between different experiments lead by different goals become difficult and meta-analyses or other forms of integration of experimental data become even more challenging. A new approach to the definition and standardization of measurement scales is needed in quantitative biology.

Measurement theory

Quantification is undoubtedly liked to progress in science, as empirical progress arguably comes from better measurements. Theoretical advances that improve the choice of what is being measured and advances in instrument design that increase the accuracy of what is being measured are two important components that drive quantitative biology (Houle et al., 2011). Most disciplines of biology in the past have focused on observational characterizations of the organisms under study, often in categorical or ordinal terms. Quantitative biology makes use of mathematical models, which make sense of the large amount of quantitative measurements obtained today (Houle et al., 2011) and is much less categorical in nature. Observations have becomes measurements and measurements have become more accurate and numerous. However, measurements must also be meaningful, particularly for the phenomenon under study. This is where modern quantitative biology lags behind more qualitative disciplines of the past.

How is the meaning of a measurement obtained? Representational measurement theory yields an explanation of how measurement and meaning arise. It determines

Scale type	Domain	Meaningful comparisons	Biological examples
Nominal	Any set of symbols	Equivalence	Species, genes
Ordinal	Ordered symbols	Order	Social dominance
Interval	Real numbers	Order, differences	Date, Malthusian fitness
Log-interval	Positive real numbers	Order, ratios	Body size
Difference	Real numbers	Order, differences	Log-transformed ratio-scale variable
Ratio	Positive real numbers	Order, ratios, differences	Length, mass, duration
Signed ratio	Real numbers	Order, ratios, differences	Signed asymmetry
Absolute	Defined	Any	Probability

TABLE 1.1: Different scale types used in the analysis of quantitative biological experiments according to Stevens, 1946. Each scale is appropriate for different types of measurements and can be meaningfully compared only to specific other scale types.

when the relation among numerical measurements assigned to attributes reflects empirical reality (Krantz et al., 1971). For example, attributes may be size or color assigned to organisms or molecular species. Measurements then consist of an assignment of numbers to attributes so that the subsequent relations among numbers can capture relations among attributes (Krantz et al., 1971). Empirical observations or comparisons then become mapped to mathematical relations and operators. There exist many different types of mathematical relations, such as order differences, ratios and equivalence relations. The correct choice depends on the hypothesis to be tested, the observed relations and the measurement process itself (Houle et al., 2011). All the steps, from the initial hypothesis to the identification of the studied entities, their attributes, the measurements and the drawn conclusions, must be well motivated and in accordance with the principles of representational measurement theory (Krantz et al., 1971). To be a valid measurement scale, a mapping must therefore be possible between the underlying empirical relational structure and the numerical relational structure. Importantly, the change of units or comparison of magnitudes should have no effect on the conclusions drawn from a particular measurement scale. These challenges are addressed in Chapter 3 for the creation of a measure of biological relevance of different test statistics. Specifically, for the integration of high-throughput data, such as transcriptomics data, an appropriate measurement scale needs to be defined when dealing with multiple features describing the a particular phenotype.

1.2 Compressed sensing

Compressed sensing is a new research field in the area of signal processing that does away with the common notion that for a signal of interest to be accurately recovered the sampling rate must be at least twice the maximum frequency present in the observed signal. This notion or limit is generally referred to as the Nyquist rate and is so fundamental that it lies at the heart of the design of most electronic devices (Candès and Wakin, 2008). However, it turns out that for a large subset of observed real world signals a more efficient sampling scheme not limited by the Nyquist rate exists, which sidesteps uniform sampling underlying the Nyquist rate limited approach. This scheme is sometimes referred to by the term compressive sampling, but is more accurately termed compressed sensing.

The underlying motivation for research in the area of compressed sensing stems from the applied signal processing techniques being overwhelmed by an increase in sensor technology and an increase in the resulting digital measurements obtained (Waters, Sankaranarayanan, and Baraniuk, 2011). However, not only the signal processing but also the storage and transmission of the acquired data has become a bottleneck (Waters, Sankaranarayanan, and Baraniuk, 2011). Therefore research in compressed sensing has become attractive, as it proposes a direct sampling of the component of the signal that is of interest, instead of completely sampling the full signal and then down-sampling it significantly after acquisition, as is commonly done for digital measurements.

The two main ingredients of compressed sensing are the sparsity assumption (Hastie, Tibshirani, and Wainwright, 2015) and the incoherence requirement (Candès and Wakin, 2008). Sparsity is a characteristic of the sampled signal, in essence the fact that the signal of interest lies on a low dimensional manifold. Many modern signals, such as image and audio recordings are of such a kind. Specifically, any observed signal that can be compressed significantly lies on a low dimensional manifold and this information can be leveraged. For example, this is the reason why high-resolution images taken with a digital camera can undergo lossy compression of several orders of magnitude and do still appear reasonably similar to the original image initially obtained, see Figure 1.8). Thus, the sparsity assumption assumes that the signal of interest lies on a low dimensional manifold. The incoherence requirement describes a more complex constraint on the sampling process itself that needs to be met and that will be discussed in the next sections. When combined, these two ingredients of compressed sensing enable efficient signal of interest recovery from a fraction of the measurements that are normally required according to the Nyquist rate limited approach.

A general drawback of compressed sensing is that while the signal of interest can

be sampled below the Nyquist rate limit, the reconstruction thereof requires significant computational power unless the signal of interest is very sparse. If the computational power to reconstruct the solution is lacking, this is in some way reminiscent finding a solution to a problem that appears under-determined. But, to overcome this challenge there have been significant advances in the mathematical theory for the design of appropriate sampling schemes and in the development of fast reconstruction algorithms for the corresponding signal recovery. These advances are described in more detail in the following sections both for the case of vector and matrix signal recovery.

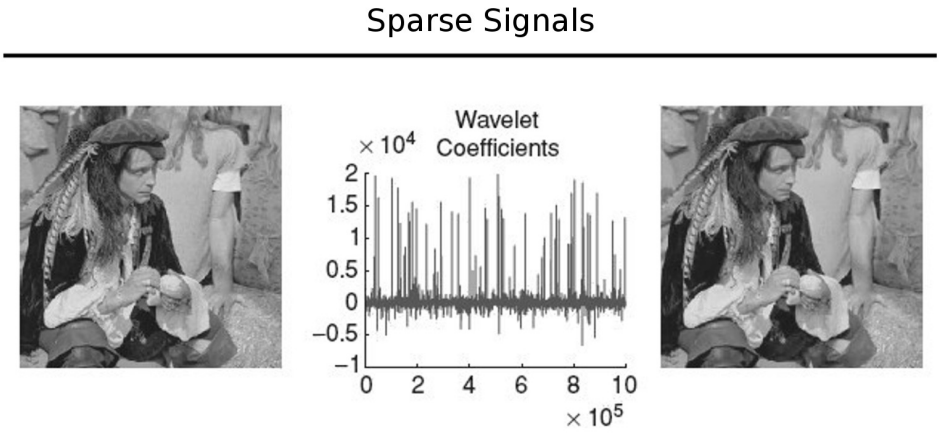


FIGURE 1.8: High-resolution image taken with a digital camera undergoing lossy compression (left, right). Both left and right image appear identical to the human eye, however the right has been significantly compressed by keeping only large wavelet coefficients. Hence, the underlying signal of interest is sparse, as is commonly the case for real world signals. The wavelet coefficients of the left image are shown in the middle. Low wavelet coefficients are dropped leading to lossy compression (right). Modified with permission from (Candès and Wakin, 2008).

1.2.1 Vector case

Sparsity assumption

Compressed sensing leverages the fact that most real world signals are sparse in some particular basis denoted by Ψ. For the vector case discussed in this section the observed signal is denoted with a lower case x. Generally this notion of sparsity can be formalized as,

$$x = \Psi\theta \tag{1.1}$$

where θ is a sparse or approximately sparse vector signal of interest. In the typical case the underlying signal θ contains few non-zero entries, as seen in Figure 1.9.

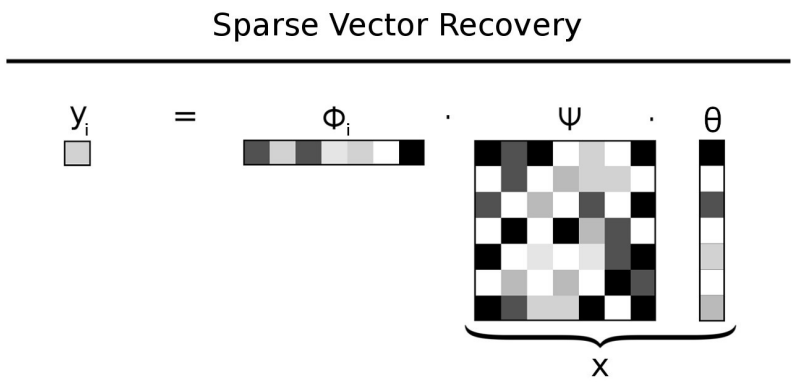


FIGURE 1.9: Compressed sensing for sparse vector recovery. For further details see Equation 1.2 and 1.3. The measurement y_i is obtained by the dot product of Φ_i the sensing basis with Ψ the canonical basis and with θ the underlying sparse signal of interest. The full vector signal x is denoted by the dot product of Ψ with θ and is what is commonly observed through uniform sampling (Nyquist rate limited). Here, dark squares indicate high values and light squares indicate low values, with white squares indicating a zero value.

Note, that the vector signal x which is normally observed may not be sparse in the canonical basis Ψ. Therefore, it might require significantly more storage space than the sparse signal θ itself and has the potential for compression. A convenient alternate basis for Ψ is a wavelet basis and by keeping only a small subset of large wavelet coefficients an almost identical signal x can be obtained that requires much less storage space in the wavelet basis Ψ. This hints at the fact that when uniform sampling is applied in the canonical basis Ψ, many of the measurements are redundant if the signal of interest θ is sparse, see also Figure 1.8 for a visual example. Thus, the sparsity assumption has important consequences for determining the sampling approaches that are feasible.

Importantly, sparsity does not only allow for signal compression, but also plays a major role in the signal acquisition process itself. It determines how efficiently signals can be acquired non-adaptively (Candès and Wakin, 2008). Typically, a vector signal x would need to be sampled completely, including all small wavelet coefficients, in order to then allow one to determine which wavelet coefficients are to be dropped and to obtain the underlying sparse signal vector θ. However, in compressed sensing a sparse vector signal θ is measured in a compressed manner directly without the computationally expensive requirement that the potentially non-sparse signal representation of x needs to be obtained or constructed, nor does the basis Ψ in which θ is sparse need to be known. In other words, an important notion in compressed sensing is that the signal can be captured efficiently, but without the need to comprehend it (Candès and Wakin, 2008). In addition, the sensing apparatus acts independently from the signal it acquires and the sparsity assumption has

a direct implication for how computationally efficient this non-adaptive process can be (Candès and Wakin, 2008).

Incoherence requirement

The incoherence requirement describes a particular constraint on the sensing process itself and is a major topic in the field of compressed sensing. It determines the restrictions on the sensing bases under which these can be used for successful signal recovery. The general sensing process or signal acquisition procedure typically follows the below scheme,

$$y = \Phi x \tag{1.2}$$

where Φ is the sensing basis, vector y denotes the obtained measurements and vector x is the observed signal. By substituting Equation 1.1 into Equation 1.2, with the observed signal composed of a sparse component θ and the canonical basis Ψ, the result can be written as,

$$y = \Phi\Psi\theta \tag{1.3}$$

where the interaction between the sensing basis Φ and the representation basis Ψ can be seen clearly. Equation 1.3 is then used in the setting of sparse signal recovery θ for the vector case, as depicted in Figure 1.9. The incoherence between basis Φ and basis Ψ describes the notion that the sampling of the sparse signal θ is done in such a way that sensing basis Φ probes the representation basis Ψ in a dense fashion, thereby exploring its space well (Candès and Wakin, 2008). More precisely, for a pair of orthobases Φ and Ψ in $\mathbb{R}^n$,

$$\mu(\Phi, \Psi) = \sqrt{n} \max_{1 \leq k,j \leq n} |\langle \phi_k, \psi_j \rangle| \tag{1.4}$$

where $\mu(\Phi, \Psi)$ is the coherence given for any two elements of the two respective bases Φ and Ψ. The basis Φ is used for sensing the vector signal x and basis Ψ is used to represent the vector signal x. Thus, according to Equation 1.4, if the bases do not contain correlated elements, the coherence is low and the incoherence is high. This condition allows for compressed sensing, for example when $\mu = 1$. The possible values of $\mu(\Phi, \Psi)$ in general fall in the range $\mu(\Phi, \Psi) \in [1, \sqrt{n}]$ (Candès and Wakin, 2008).

A particular pair of bases that have low coherence and are thus sufficiently incoherent are the wavelet basis Ψ paired with the noiselet basis Φ (Coifman, Geshwind, and Meyer, 2001; Candès and Wakin, 2008). For the particular coherence between wavelets and noiselets $\mu(\Phi, \Psi) = \sqrt{2}$ (Candès and Wakin, 2008). More extreme cases of incoherence include the canonical basis and the Fourier basis, with $\mu(\Phi, \Psi) = 1$

and thus maximal incoherence (Candès and Wakin, 2008). Another simple to generate case and one which plays a major role in theoretical proofs of compressed sensing are random orthobases for sensing basis Φ, which yield with high probability $\mu(\Phi, \Psi) = \sqrt{2 \log n}$ for any Ψ (Candès and Wakin, 2008). This is important, as often the underlying basis Ψ in which the signal of interest θ is sparse is not known. Random orthobases extend to any sub-Gaussian matrix (Rivasplata, 2012), which includes random Gaussian or Rademacher matrices. Thus, as incoherence is beneficial for efficient sampling, random orthobases denote an efficient mechanism at signal acquisition (Candès and Wakin, 2008). Notably, some of the above depicted bases have important computational advantages leveraged in compressed sensing, besides fulfilling the incoherence requirement. Further details of sensing bases, which are termed measurement operators in the setting of signal recovery, are discussed in the next sections.

Signal of interest recovery

In the framework of compressed sensing recovery of a sparse vector signal θ requires the solving of an under-determined system of linear equations, see Equation 1.3 and Figure 1.8. Unfortunately, the inverse problem of Equation 1.3 that needs to be solved is generally NP-hard (Waters, Sankaranarayanan, and Baraniuk, 2011). Furthermore, if there are not sufficient measurements y available the problem is under-determined. A major result from recent theoretical advances in compressed sensing is that such recovery is still possible despite these hurdles and also with relatively efficient algorithms, if one can make the assumption that the vector signal x is indeed sparse, e.g. contains a sparse signal of interest θ in some basis Ψ and measurements y are sufficient for this sparsity. These notions will be made more precise shortly in the next sections.

During the initial development of the field of compressed sensing a major research focus was on theoretical advances in the recovery of sparse vectors, as is the focus in this section. Only later was the process generalized to the matrix case discussed in Section 1.2.2. Recovery in the case of an observed vector signal x is typically performed though convex optimization or greedy iteration. See the following section on reconstruction algorithms for a description of the different optimization routines. The optimization problem is given as,

$$\min ||\theta||_0 \quad \text{subject to} \quad y = \mathcal{A}(x), \tag{1.5}$$

where the measurements are denoted by y, the observed signal vector is denoted as x, being composed of canonical basis Ψ and sparse signal vector θ, and is per sparsity assumption sparse in some transform domain. $||\theta||_0$ denotes the number of non-zero entries in θ that are to be minimized. $\mathcal{A}$ denotes an under-determined linear operator (Waters, Sankaranarayanan, and Baraniuk, 2011) with the property

that $\mathcal{A} : \mathbb{R}^n \rightarrow \mathbb{R}^p$ and $p \ll n$. In case of a convex recovery the l_0 norm is relaxed to the l_1 norm and if the number of measurements m obtained uniformly at random in the Φ domain are on the order of,

$$m \leq C \cdot \mu^2(\Phi, \Psi) \cdot S \cdot \log n \tag{1.6}$$

where C is a positive constant and S is an integer describing the sparsity of an S-sparse vector signal, then the probability of recovering the signal of interest θ with a convex optimization program is exceedingly high (Candès and Wakin, 2008). However, some further restrictions exist and are described by Candes and Romberg, 2007 in further detail for this specific idealized case. Importantly, real world signals are typically not exactly sparse and contain measurement noise. This makes recovery more challenging and only feasible if the following additional property is fulfilled by the measurement operator,

$$(1 - \delta_K) \left\|\theta\right\|_2^2 \leq \left\|\mathcal{A}(x)\right\|_2^2 \leq (1 + \delta_K) \left\|\theta\right\|_2^2, \qquad \forall ||\theta||_0 \leq K \tag{1.7}$$

This denotes the restricted isometry property (RIP) and is a cornerstone of compressed sensing when performed on real world measurements with a certain degree of measurement noise. Thus, given the incoherence requirement and sparsity assumption discussed in Section 1.2.1, signal recovery can be performed as denoted in Equation 1.5.

Measurement operators

When the underlying sensing basis Ψ in which the to be recovered signal θ is sparse is not known, as is frequently the case, the choice of an incoherent sampling basis Φ is challenging. The sampling basis Φ is termed measurement operator in this context and in addition to the incoherence requirement needs to satisfy the RIP condition, as discussed in the previous section on signal of interest recovery, to enable efficient recovery in a real world setting. Random matrices offer a solution when the underlying basis Ψ is not known and have been widely used and studied in the field of compressed sensing to address this problem, as noted in the previous section on the incoherence requirement. According to Candès and Wakin, 2008 random matrix based measurement operators that can operate without knowledge of the underlying basis Ψ can be constructed by at least four different methods,

1. Sampling n column vectors uniformly at random on the unit sphere $\mathbb{R}^m$.
2. Sampling i.i.d. entries from a normal distribution with $\mu = 0$ and $\sigma = 1/m$.
3. Sampling i.i.d. entries from any other sub-gaussian distribution.
4. Sampling a random projection P and normalizing with $\sqrt{n/m}\, P$.

where n is the size of the signal and m are the number of measurements. With high probability all the above constructions of basis Φ yield measurement operators which satisfy the RIP condition, given that there exist enough measurements m for signal of interest recovery according to,

$$m \geq C \cdot S \log n/S \tag{1.8}$$

where C is a constant and S is an integer describing the sparsity of an S-sparse vector signal. Thus, such random matrix based measurement operators can be considered *universal* in the sense that these can be designed without knowledge of the to be recovered signal basis Ψ.

Furthermore, there exist dense and sparse measurement operators that are sparse akin to the sparse signal of interest θ. For the vector case, it was shown that very sparse measurement operators in the form of random projections, such as sub-Gaussian measurement operators with only a small fraction of non-zero entries, can accurately recover an underlying signal (Li, Hastie, and Church, 2006). Notably, the benefit of utilizing sparse measurement operators is the reduced recovery time required (Berinde and Indyk, 2008; Li and Zhang, 2015) and reduced storage space needed (Cai and Zhang, 2015), which are both a current bottleneck of compressed sensing based approaches to signal recovery in real world settings.

Reconstruction algorithms

The recovery algorithms applied in practice are based on convex optimization or greedy iteration of the under-determined optimization problem defined in Equation 1.5, where the measurements were denoted by y, the observed signal vector was denoted as x, $||\theta||_0$ denoted the number of non-zero entries in signal of interest θ and the operator $\mathcal{A}$ was a linear operator (Waters, Sankaranarayanan, and Baraniuk, 2011) with the property that $\mathcal{A} : \mathbb{R}^n \rightarrow \mathbb{R}^p$ with $p \ll n$. The optimization problem is thus under-determined. In practice, linear operator $\mathcal{A}(x)$ is simply a dot product sensing basis or measurement operator Φ with the observed signal x.

A specific approach at sparse signal of interest recovery is based on convex optimization and is termed basis pursuit (Chen, Donoho, and Saunders, 2001). In the case of basis pursuit, Equation 1.5 is smoothed by replacing the l_0 norm with an l_1 norm, resulting in the modified optimization problem with threshold ϵ,

$$\min ||\theta||_1 \quad \text{subject to} \quad ||\Phi x - y||_2^2 \leq \epsilon \tag{1.9}$$

This convex optimization problem of Equation 1.9 can be recast as a linear optimization program such as quadratic programming, which in turn can utilize very efficient general purpose solvers. These solvers include interior point methods (Potra and Wright, 2000), sequential shrinkage (Ghosh, Nickerson, and Sen, 1987) and iterative

shrinkage methods (Beck and Teboulle, 2009). The original NP-hard optimization problem becomes thus tractable in practice (Candès and Wakin, 2008; Candes and Romberg, 2007).

Another approach at solving sparse signal of interest recovery is based on the greedy method. Specifically, the matching pursuit algorithm has been developed to tackle this challenge (Mallat and Zhang, 1993). Equation 1.5 can be seen as simply a backwards progression. Each element of θ_i is estimated based on all measurements y in combination with the particular column in Φ which results in the largest absolute value $|\theta_i|$. The progression continues for all elements θ_i in θ and iterates until the error is below a specified threshold, ϵ. A more advanced approach termed orthogonal matching pursuit (OMP) (Tropp and Gilbert, 2007) uses a re-evaluation of all θ_i by least squares after each iteration of sampling all columns of Φ. Thus, this later approach is less affected by inadequate starting conditions. Further improvements termed StOMP and CoSaMP are discussed by Donoho et al., 2012 and Needell and Tropp, 2009, respectively.

1.2.2 Matrix case

Recently many of the algorithms and theoretical advances in the field of compressed sensing have been extended from the vector case to the matrix case. The signal recovery problem for the matrix case is commonly termed matrix recovery or low-rank/affine matrix recovery and will be discussed in this section with a focus on the differences to the vector case. Matrix recovery is an important concept particularly relevant for the algorithm developed in Chapter 2.

Sparsity assumption

As mentioned in Section 1.2.1, sparsity is a characteristic of the sampled signal. The sparsity assumption in the setting of matrix recovery is equally based on the idea of compressibility. For the vector case, a vector signal x is compressible if it only contains few non-zero elements in an appropriate basis Ψ, see Section 1.2.1. A similar notion is applied in the matrix case. However, instead of few non-zero elements in the matrix X, few non-zero elements in the diagonal matrix of singular values d of the matrix X are referred to by what is mean with sparse, as seen in Figure 1.10 where only the non-zero elements of matrix d are shown. A matrix with few non-zero entries in its diagonal is commonly termed a low-rank matrix and is an example of a low dimensional manifold and also a sparse matrix. Notably, a low-rank matrix can approximate any more complex matrix signal to a specified degree. Thus low-rank matrix recovery allows the approximate recovery of any matrix signal to a specified degree of sparsity, which is important for real world applications.

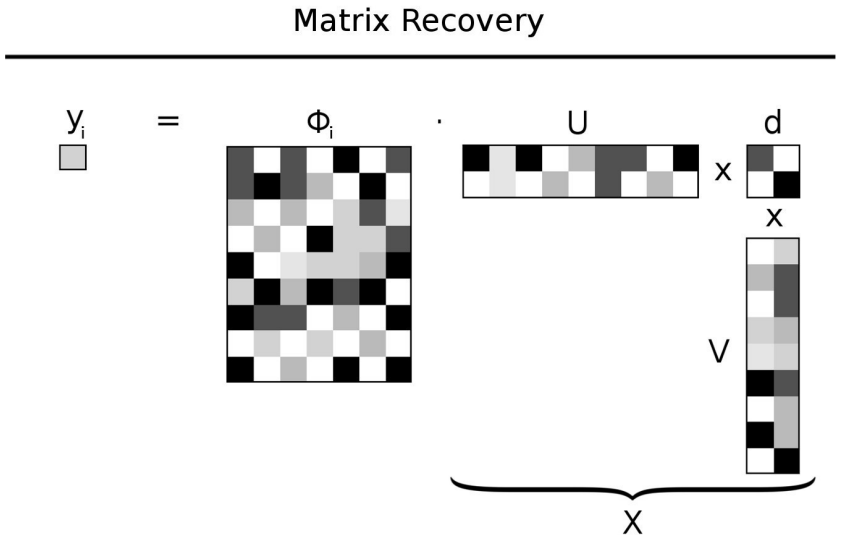


FIGURE 1.10: Compressed sensing for sparse matrix recovery. The measurement y_i is obtained by the dot product of Φ_i with the left singular vector U, the singular value containing matrix d with few non-zero entries on the diagonal, and the right singular vector V. The full signal matrix X is denoted by the above singular value decomposition of UdV, which stands for a sparse signal, as the dimensionality of matrix d is much less than the of X.

In matrix recovery the matrix signal of interest X can be defined as sparse according to the singular value decomposition,

$$X = UdV \tag{1.10}$$

where d is a sparse or approximately sparse diagonal matrix, U is the left singular vector matrix and V is the right singular vector matrix. The full signal X is denoted by the above singular value decomposition of UdV. Thus, the full signal X does not necessarily contain few non-zero elements, only its diagonal contains few non-zero elements, depending on its degree of sparsity, see Figure 1.10. If a matrix is sparse in this way, matrix recovery may be feasible, if in addition an incoherence requirement is satisfied similar to the vector case.

Incoherence requirement

Matrix signals in the real world are typically not exactly sparse and contain measurement noise, which makes recovery more challenging, as described for the vector case. Specifically, the incoherence μ as defined in Section 1.2.1 is used to determine a type of RIP condition that in addition to sparsity is necessary for signal recovery. While the RIP condition was developed for the vector case in Section 1.2.1 it also applies to the matrix case if extended accordingly. Specifically, for the sparse diagonal matrix d, the RIP condition can be applied directly if the diagonal itself is treated as

a vector. In general, for the matrix case the cardinality of a vector is replaced by the rank of the matrix and the Frobenius norm is replaced by the Euclidean norm. Thus for the matrix case the RIP condition is termed Rank Restricted Isometry Property (RRIP) and is accordingly defined as,

$$(1-\delta_r)\,\|X\|_F^2 \leq \|\mathcal{A}(X)\|_F^2 \leq (1+\delta_r)\,\|X\|_F^2\,, \qquad \forall\,\mathrm{rank}(X) \leq r \tag{1.11}$$

where X is a low-rank matrix and $\mathcal{A}$ takes the form of a measurement operator with the property that $\mathcal{A} : \mathbb{R}^{n\times m} \rightarrow \mathbb{R}^p$ with $p \ll nm$. The incoherence condition for a specific case is defined by constant δ_r (Recht, Fazel, and Parrilo, 2010).

Signal of interest recovery

For the scenario of affine or low-rank matrix recovery the optimization problem is conceptualized as,

$$\min \mathrm{rank}(X) \quad \text{subject to} \quad y = \mathcal{A}(X), \tag{1.12}$$

where X is a low-rank matrix and $\mathcal{A}$ takes the form of a measurement operator with the property that $\mathcal{A} : \mathbb{R}^{n\times m} \rightarrow \mathbb{R}^p$ with $p \ll nm$. In case of a convex relaxation akin to the vector case in Section 1.2.1, the rank operator is relaxed to the nuclear norm. Then, the signal $X \in \mathbb{R}^{n\times m}$ with $\mathrm{rank}\,(X) \leq r$ can be efficiently recovered from a small set of measurements $\mathbf{y} \in \mathbb{R}^p$ with a given measurement operator if the incoherence condition of Section 1.2.2 is satisfied accordingly. The recovery problem can then be pursued as follows,

$$y = \mathcal{A}(X) = \langle X, A_i \rangle, \quad i = 1, \ldots, q. \tag{1.13}$$

where y is a vector of the q measurements which are a linear combination of the true matrix signal X and the sensing matrices A_i. If the matrix signal X is to be recovered from q measurements the problem is under-determined and there exist an infinite number of matrix signals X' which satisfy $y = AX'$ if q is less than the number of required measurements (Candès and Wakin, 2008). However, if the matrix signal can be assumed to be sparse, as defined in the previous sections, the number of candidates for X' is reduced drastically. Thus, if the measurement matrix is incoherent and the signal sparse the recovery is possible and efficient as for the vector case.

Measurement operators

In the case of matrix recovery the standard measurement operators from the vector case apply and of these especially random matrices are frequently used. However, with respect the algorithm developed in Chapter 2, it is informative to discuss which other measurement operators are feasible. There exists only limited literature

on the design of non-standard measurement operators for the matrix case. A motivation for further such research is the immense storage space require for computations with certain measurement operators. One particular direction of research is concerned with measurement operators which are rank-1 projections or derivatives thereof (Cai and Zhang, 2015). These types of measurement operators are significantly more efficient to store while still allowing accurate signal recovery. For example, the sub-Gaussian based random matrix design introduced for the vector case requires a minimum of $\mathcal{O}(m \cdot n \cdot k)$ bytes of storage space, where m is the number of measurements, n is the size of the row dimension and k is the size of the column dimension. This means that a common 10,000 $\times$ 10,000 measurement matrix of rank 10 requires 45TB of storage space, some of it potentially in memory (Cai and Zhang, 2015). This is prohibitively large for most applications. Another approach uses very sparse random projections to achieve a similar goal of reducing storage space (Li, Hastie, and Church, 2006). However, compared to the design of measurement operators for the vector case, the matrix case is much less well studied.

Reconstruction algorithms

Current reconstruction algorithms for the case of matrix recovery can be categorized into convex and greedy algorithms. Convex algorithms are based on a nuclear norm regularization, as detailed in Section 1.2.2. Thus, the convex optimization problem can be recast as a semidefinite programming problem, akin to a linear optimization program for the vector case. Subsequently, efficient solvers can be utilized, which include SeDuMi (Sturm, 1999), SDPT3 (Toh, Todd, and Tütüncü, 1999) and SDPA (Fujisawa et al., 2000). However, the later algorithms typically only work on small scale problems that are uninteresting in a real world setting.

Another strategy are greedy algorithms, which are typically based on an iterative singular value thresholding approach initially developed by (Cai, Candès, and Shen, 2010). Specific improvements include Iterative Hard Thresholding (Blumensath and Davies, 2009), Normalized Iterative Hard Thresholding (Blumensath and Davies, 2010) and Conjugate Gradient Iterative Hard Thresholding (Blanchard, Tanner, and Wei, 2015). An additional greedy algorithm is Atomic Decomposition for Minimum Rank Approximation (ADMiRA) (Pilastri and Tavares, 2016) and yet another such heuristic is based upon the constrained search space of low-rank matrices that uses Riemannian optimization (Pilastri and Tavares, 2016). The later approach is used for the algorithm developed in Chapter 2, as it exploits the smooth geometry of fixed-rank matrices which has been shown to be more scalable than other approaches and allows for a particularly broad scope in matrix recovery (Tan et al., 2014).

1.2.3 Applications

The applications of compressed sensing are numerous (Candès and Wakin, 2008). On one hand, the idea of using a random measurement operator to recover a sparse signal can be seen as an alternative approach to data compression. According to Candès and Wakin, 2008 such a randomly designed measurement operator Φ can be seen as a universal encoding strategy, since it is independent of the representation basis Ψ. This type of compression seems to be particularly useful for image acquisition and sensor networks, where there are multiple signals from multiple sources (Candès and Wakin, 2008) and bandwidth is limited. On the other hand, another application involves the design of error correcting codes, especially with the aim of controlling for more systematic sources of noise and a speed-up in the computational efficiency of channel coding (Candès and Wakin, 2008). Lastly, compressed sensing may enable the solving of inverse problems, such as blind deconvolution and self-calibration. If the measurement operator Φ is given or predetermined by the observed signal and a basis Ψ also exists, which is additionally sufficiently incoherent with Φ, then such blind recovery becomes possible. This may lead to improvements in the correction of systematic biases in the observed signal, as will be discussed in Chapter 2.

Acquisition

There is likely to be a profound interaction between software based approaches to compressed sensing and the development of compressive sensing hardware (Candès and Wakin, 2008). Candès and Wakin, 2008 propose scenarios where physical sampling devices that record directly the discrete, low-rate incoherent measurements of analog signals may outperform conventional hardware based on CCD and CMOS technologies. Especially in the area of imaging this seems promising, since conventional technologies are limited to the visible spectrum and slow sampling rates on the order of GHz (Candès and Wakin, 2008). The hardware for compressed sensing is currently based on digital micro-mirror devices (DMDs), which allow the construction of random test functions by alternating mirror positions. Thus, each measurement is a convolution of a test function with the scene of interest, from which an image is then reconstructed (Duarte et al., 2008). In other words, a random linear measurements of the scene of interest is obtained. For the infrared spectrum and other spectra outside the visible range, such an approach is deemed to be more cost effective than current technologies (Duarte et al., 2008). Also, compressive measurement sensors can potentially be constructed on a much smaller scale then current imaging hardware. The main components of single pixel cameras are currently a DMD, two lenses, a single photon detector and an analog-to-digital converter (Duarte et al., 2008). The ability to then directly measure a compressed image enables the handling of large data streams of video or hyper-spectral images obtained by satellites.

A recent focus is on the application of algorithms, such as classification, directly on compressive measurements obtained by these sensors, without requiring full reconstruction of the underlying signal (Davenport et al., 2007).

Calibration

A particular application of compressed sensing that is related to the algorithm presented in this thesis is the use of compressed sensing in blind deconvolution. In the framework of an inverse problem this can be stated mathematically as,

$$y = \mathcal{A}(h, X) \tag{1.14}$$

where y is a vector of known measurements, $\mathcal{A}$ is a linear operator that is dependent on h and X is the matrix signal observed. Here, h needs to be estimated as well, making the problem in Equation 1.14 a challenge compared to the simpler Equation 1.12. Typically, prior information about the measurement setup or additional sparsity assumptions can be used to make this problem identifiable and such an approach is common in the setting of self-calibration (Ling and Strohmer, 2015). As the design of ever more precise sensing devices becomes increasingly difficult, the need for precise calibration of such devices increases as well (Ling and Strohmer, 2015). Calibration is necessary for the optimal performance of the developed devices, which can include everything from micro-scale sensors to devices powering large-scale telescopes (Ling and Strohmer, 2015). Ideally, the sensing devices only take measurements and do not require self calibration, which would simply be performed *post hoc* by solving the above mathematical problem. This approach is often prohibitively expensive in computational terms (Ling and Strohmer, 2015). However, by the use of lifting techniques (Ahmed, Recht, and Romberg, 2014), inverse problems can potentially be solved efficiently. But, this generally requires the assumption of sparsity for the signal and for the systematic error or calibration model. Then, the above formulated problem (Equation 1.14) can be solved via convex programming (Ahmed, Recht, and Romberg, 2014) and potentially leverage the recovery algorithms described in Section 1.2.2.

Chapter 2

Blind compressive normalization (BCN)

2.1 Introduction

The rise of high-throughput technologies in the domain of molecular and cell biology, as well as medicine, has generated an unprecedented amount of quantitative high-dimensional data. Public databases at present make a wealth of this data available, but for meaningful analyses integrating different experiments and technologies appropriate normalization is critical. Without such appropriate normalization, meta-analyses can be difficult to perform and the potential to address shortcomings in experimental designs, such as inadequate replicates or controls, via the reuse of public data is limited.

The overarching problem for data integration is that of normalization, which is becoming more apparent and limiting as the need for reuse and re-analysis of high-throughput data rises. Normalization involves the attenuation of bias resulting from confounding factors affecting the measurement process. Technical bias of an instrument or sample preparation procedure can be addressed by measuring identically processed standards of known quantity. Use of such standards is widespread in serial technologies. The further up-stream in the measurement process quantitative standards are introduced, the more potential sources of bias can be accounted for. Biological bias due to non-identical cells or organisms is often addressed instead by randomization (Montgomery, 2008). This later approach presupposes that the contrast of interest and potential bias sources are known. An overview of potential bias sources with a focus on high-throughput technologies is given by Lazar et al., 2012.

High-throughput technologies are challenging to normalize especially because the bias of biological molecules measured in parallel is not independent. Such non-independent bias stems from molecular interactions throughout the measurement process, including sample preparation procedures and instrument settings that are

dependent on the measured sample itself and its biological signal. Quantitative measurement standards must therefore effectively cover a vast number of possible combinations of potential signals measured. In addition, instrument or measurement process components are sometimes one-time-use, such as in the case of microarray technologies, making appropriate normalization with measurement standards unfeasible. In part for these reasons, high-throughput technologies have been designed with a focus on relative comparisons, such as fold changes, rather than absolute quantification. While a limited number of spike-in standards can account for some technical bias (Lovén et al., 2012) sample preparation procedures that are important sources of bias, such as library preparation, protein extraction or metabolic labeling, generally happen up-stream of spike-in addition. Bias attenuation by randomization is not generally possible, as contrasts of interest are not initially known in the exploratory analyses typically performed with high-throughput technologies.

Initial experimental design establishes how quantitative measurement standards or randomization are employed in a particular experiment. However, in the case of experiments that draw on samples from public databases, the attenuation of bias must be attained *post hoc*. Normalization methods that can be applied across large scale public databases for the purpose of normalization are currently limited to approaches that demand ad hoc assumptions about noise sources and the biological signal. This is due to a lack of quantitative standards and insufficient or incoherent experimental annotation available in public databases. Quantitative standards would allow a straightforward correction of bias and sufficient annotation would allow supervised normalization methods to be applied on a large scale (given that appropriate replicate experiments have been performed). But, this is not generally the case in the publicly available experiments. Hence, current approaches must make assumptions not necessarily appropriate for all the experiments contained in a public database in order to proceed with normalization. In general, the appropriateness of such assumptions is challenging to evaluate. Simple averaging and scaling techniques that do not rely on explicit assumptions are computationally feasible when dealing with large high-throughput databases, but do not take advantage of the large amount of public high-throughput data currently available as prior knowledge. A more systematic approach to normalization is required that leverages the available public data without requiring prior knowledge about the specific experimental conditions or the general availability of quantitative standards.

In the following section, the state of the field is reviewed with respect to high-throughput data normalization techniques. The focus is placed on the progression from unsupervised to supervised normalization techniques. Next, the developed algorithm is described in detail. It leverages detectable redundancies in public high-throughput databases, such as related samples and features to perform blind normalization of common but unknown bias sources. Highlighted is an evaluation of

its performance and robustness in simulation depicted in Figure 2.3 to Figure 2.10. Different variants of the proposed algorithm exist that in the framework of compressed sensing progress from *entry* sensing towards k-sparse sensing and subsequently blind recovery. The underlying assumptions of the algorithm are discussed in Section 2.3.3 and biological validation experiments are proposed in Section 2.3.5. In order to provide researchers an efficient way to apply the developed algorithm a software package has been created[1]. More specific details regarding the implementation are given in Supplementary Material A.

2.2 State of the field

The normalization methods reviewed in this section are distinguishable on the level of the assumptions made with respect to bias sources or the biological signal. The focus of the surveyed normalization methods lies generally not on computational efficiency, but on attenuating different bias sources with prior information obtained from the experimental design or a given setting. This is because most normalization techniques have not been developed for large scale meta-analyses or normalization of entire databases and the performance has thus only played a minor role. As an ideal and simple normalization approach, standardization with quantitative standards could be applied, if such measurement standards were established and feasible to obtain for all high-throughput experiments contained in current public databases. In this approach then, the measurement processes could be effectively standardized before the particular high-throughput data is obtained. However, quantitative standards are currently unfeasible to obtain due to various reasons (see Section 2.3) even tough this would be an appropriate and effective approach to normalization. The available methods of normalization that do not rely on quantitative standards are reviewed below, from unsupervised to supervised learning (see Figure 1.6).

Unsupervised

Unsupervised approaches generally make use of ad hoc assumptions about noise sources or the biological signal, which are then leveraged in an attempt to average out bias in a *post hoc* fashion. The earliest methods developed in this area are concerned with centering and scaling (Cheadle et al., 2003). Notably, some methods assume that an appropriate scaling is obtained by variance stabilization across features (Huber et al., 2002). Other methods deemed appropriate are scaling across samples based on the assumption that the overall biological signal does not vary significantly across samples. The later method is known as quantile normalization (Bolstad et al., 2003) and still widely used for normalization as a component of the

[1] http://github.com/a378ec99/bcn

RMA method (Irizarry et al., 2003). More bias source specific methods assume that the bias inherent in the detection of features is the most important to address, such as GC effects in transcriptomics measurements (Wu and Irizarry, 2004; Binder and Preibisch, 2008; Piccolo et al., 2012). Other methods yet, focus on recovering artifacts introduced by the measurement instrument (Moffitt et al., 2011; Suárez-Fariñas et al., 2005). For example, Shabalin et al., 2008 are concerned with normalization between different instrument components, such as microarray platforms, which is a critical concern for the meta-analyses discussed here. However, currently all unsupervised approaches fail to exploit the wealth of information that is available in public high-throughput databases. In addition, it is difficult to assess the appropriateness of particular assumptions about noise sources and biological signal that are exploited with any particular method covered here.

Supervised

Supervised approaches generally make use of replicate samples or prior knowledge of potential confounding factors and contrasts of interest, in order to then perform a form of randomization to average out bias. Ideally, if contrasts of interest have replicates overlapping with known confounding factors, these can subsequently be leveraged to remove bias in feature or sample space, for example via simple linear centering (Li and Wong, 2001) or more complex non-linear adjustments (Benito et al., 2004). For small sample sizes (<25 samples) the popular empirical Bayes method ComBat has been designed (Johnson, Li, and Rabinovic, 2007) based on this strategy. However, ComBat and derivative methods are unable to detect and remove bias outside of replicate samples that have been specifically designed in the experimental setup to limit known confounding factors. If the available replicates are insufficient, but the contrast of interest is instead known *a priori*, then the Surrogate Variable Analysis (SVA) method by Leek et al., 2012 can be applied. Alternatively, a combination thereof with centering and scaling (Hornung, Boulesteix, and Causeur, 2016) is also applicable. However, supervised approaches are generally unfeasible for large public databases due to limited annotation and unknown contrasts of interest, which are required for supervised normalization.

Semi-supervised

Semi-supervised approaches have been introduced recently. These implicitly or explicitly aim to exploit additional data to learn underlying parameters that can be transferred to the dataset at hand. In particular, fSVA (Parker, Bravo, and Leek, 2014), fRMA (McCall et al., 2011) and the Gene Expression Commons (Seita et al., 2012) take such an approach. The later two methods aim to adjust the weight or scale parameters of individual features based on global distributions obtained from

the additional data. The fSVA method requires knowledge of contrasts of interest for the additional data to be of use and therefore is impractical in the case of exploratory analyses. Lastly, the XPN method (Shabalin et al., 2008) is a tangential approach to the attenuation of bias. It applies a smoothing based on unsupervised clustering of the additional data and uses information from both features and samples space. It may be used in combination with other unsupervised or supervised normalization methods. However, XPN is specific for the case of normalizing between two particular datasets and correcting for the bias across one particular contrast only. Overall, the exploitation of additional data from high-throughput databases for normalization only remains to become more prominent and effective as database sizes increase.

2.3 Algorithm

The following section presents a systematic approach to the *post hoc* removal of bias in high-throughput databases as outlined in Figure 2.1. The proposed approach is formulated in the theoretical framework of compressed sensing and uses scalable Riemannian optimization for the recovery of bias. Importantly, as database sizes increase, more complex bias can be normalized. In addition, the approach accounts for missing values and can incorporate side information such as spike-ins. Most of the material highlighted in the following sections is taken verbatim or in modified form from publication A.1 (Ohse, Börries, and Busch, 2019).

The challenge of normalizing large high-throughput databases is distinct from the traditional $p \gg n$ problem (Friedman, Hastie, and Tibshirani, 2001) of high-throughput data normalization, since the samples (n) accumulated in the public domain have increased substantially (see Figure 1.5) over the number of features (p) that have stayed relatively constant. More importantly, the current objective is not the analysis of single samples but the analysis of a large numbers of samples at the same time, such as entire high-throughput databases. Both the feature and sample space of databases are to date large and on approximately the same order of magnitude, with approximately 100,000 samples by 100,000 features for transcriptomics data. If lower level features, such as probes or reads are considered, the number of features increases by a factors of 10-100. Therefore, computational scalability becomes an important consideration when trading-off accuracy and feasibility. Recent advances in the area of machine learning based on the sparsity assumption (Hastie, Tibshirani, and Wainwright, 2015) have shown that limited sampling of high-dimensional signals is often sufficient for efficient recovery. For example, in the area of collaborative filtering, large low-rank matrices are routinely recovered from a small number of sampled entries (e.g. *entry* sensing) (Mazumder, Hastie, and Tibshirani, 2010; Jain, Netrapalli, and Sanghavi, 2013; Vandereycken, 2013). If confounding

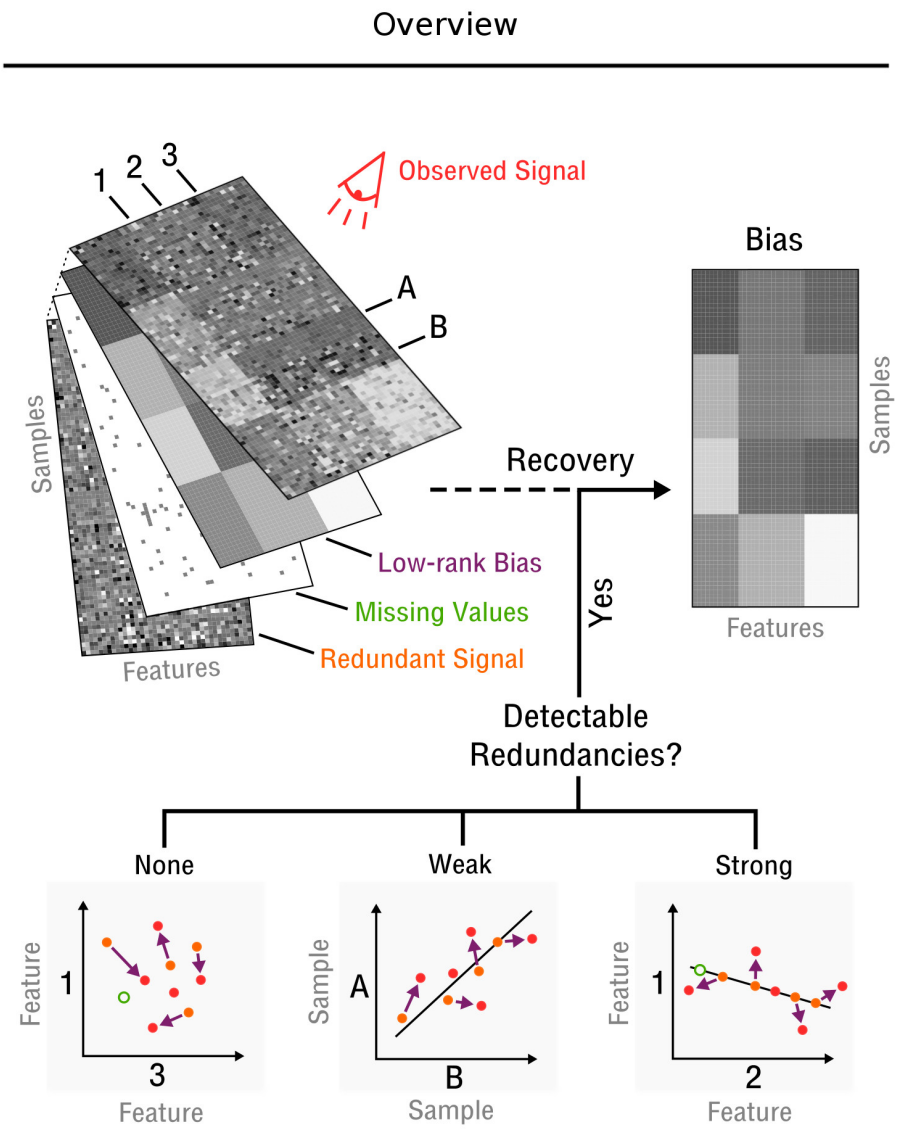


FIGURE 2.1: Blind recovery of bias in high-throughput databases (Ohse, Börries, and Busch, 2019). (Top) A database consisting of features (e.g. measurements of RNA, protein or metabolites) and samples (e.g. different cell types under various stimuli). Recovery of the underlying bias is feasible, if some redundant signal exists that is incoherent with the bias and is partially detectable from the observations. (Bottom) Redundancies characterized as detectable and as weak or strong based on the dependency strength between features or samples. The more a redundant signal (orange dots) falls on the curve (black line) the stronger is the redundancy.

factors in high-throughput databases are equally amenable to the sparsity assumption, bias due to the measurement process may be recovered from a relatively small number of sampled entries in the form of measured quantitative standards. However, since such standards are not available or feasible to obtain *post hoc* for current

high-throughput databases, its is proposed to instead utilize database wide redundancies for bias recovery. This approach uses similar optimization routines and recovery theory as approaches based on sampled entries (e.g. *entry* sensing) once the observed redundancies are transformed into the correct framework.

By assuming that confounding factors are sparse, the proposed problem of bias recovery becomes manageable via efficient manifold optimization techniques (Vandereycken, 2013). The larger a high-throughput database becomes the more effectively one can bootstrap off of database wide redundancies for a particular level of sparsity in the bias. This scaling effect is critical, as blind bias recovery requires a certain number of detectable redundancies in the form of accurately estimated dependencies to effectively recover bias. Thus, for large high-throughput databases the proposed bias recovery process becomes more effective. The main innovation of the algorithm is the casting of detectable redundancies in high-throughput databases as prior knowledge of the measurement operator design. In the framework of compressed sensing this enables blind recovery of bias and subsequent normalization of high-throughput databases from merely estimated dependencies. Thereby, more restrictive assumptions on the biological signal or noise sources common in other unsupervised normalization approaches are sidestepped. Additional normalization approaches can still be applied as pre-processing if so desired before the proposed normalization algorithm is applied.

For the biological or medical researcher working with high-throughput data this means that when blind compressive normalization is applied to a database which include their samples, these samples are made more comparable to each other and overall to other samples in the database, as bias stemming from unknown confounding factors is attenuated.

The developed bias recovery algorithm and subsequent blind normalization of high-throughput data begins with a general assumption that there are a limited number of confounding factors that markedly affect the measurement process (sparsity assumption). Hence, bias is modeled as a low-dimensional manifold that in this particular case takes the form of a low-rank matrix $\mathbf{X}$. Low-rank matrices are a flexible model, that can approximate arbitrarily close any true underlying bias. For example, a $\mathrm{rank} = 10$ approximation of an image generally appears indistinguishable to the full rank image to human perception. However, other low-dimensional manifolds, such as symmetric or spherical manifolds, are equally applicable and can be used with the proposed algorithm.

In the framework of compressed sensing the resulting matrix recovery problem that aims to recover an underlying low-rank bias matrix $\mathbf{X}$ is defined as follows (Tan

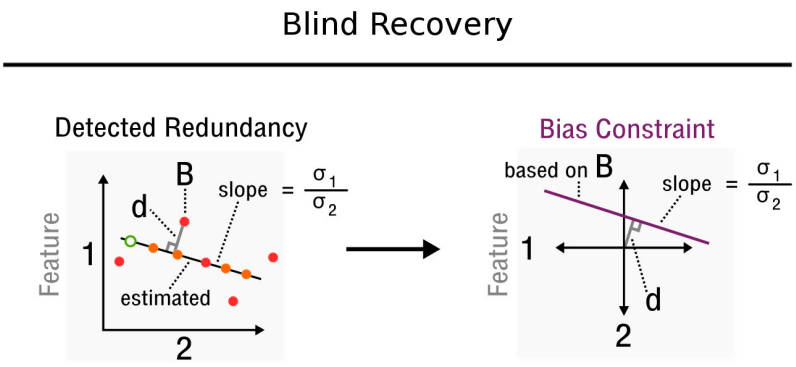


FIGURE 2.2: The measurement inference process from detected redundancies to bias constraints. (Left) In feature space a redundancy is detected (black solid line) and sample $\mathbf{B}$ allows the characterization of distance $\mathbf{d}$ and slope $\frac{\sigma_1}{\sigma_2}$. (Right) The corresponding bias based on sample $\mathbf{B}$ is denoted in this new feature space, where $\mathbf{d}$ characterizes the offset from the origin and all bias estimates must lie on the given curve for zero error (solid purple line).

et al., 2014).

Definition 1. Given a linear operator $\mathcal{A} : \mathbb{R}^{n\times m} \rightarrow \mathbb{R}^p$, let $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \epsilon$ be a set of p measurements of an unknown rank $\hat{r}$ matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ and noise ϵ. Matrix recovery solves the problem of $\min_{\mathbf{X}} \|\mathbf{y} - \mathcal{A}(\mathbf{X})\|_2^2$ subject to $\text{rank}\,(\mathbf{X}) \leq r$, where $p \ll nm$ and $r \geq \hat{r}$.

The specific type of linear operator used in Definition 1 depends on the context and is commonly defined as the Frobenius inner product of matrix $\mathbf{X}$ and sensing matrices $\{\mathbf{A}_i \in \mathbb{R}^{n\times m}\}_{i=1,\dots,p}$ such that $\mathbf{y}_i = \sum_{j=1}^{n}\sum_{k=1}^{m}(\mathbf{A}_i)_{jk}\mathbf{X}_{jk}$. In the general case of *dense* sensing, sensing matrices $\mathbf{A}_i$ are defined $\forall\ j \in \{1,\dots,n\}$ and $\forall\ k \in \{1,\dots,m\}$ as $(\mathbf{A}_i)_{jk} \sim \mathcal{N}$. Notably, this approach at bias recovery presupposes a measurement setup that provides prior information about $\mathbf{A}_i$ and $\mathbf{y}_i$ to recover matrix $\mathbf{X}$ according to Definition 1. It is shown here that such prior information can be indirectly obtained from an approximation of the redundancies that commonly exists in high-throughput databases (see Section 2.3.1) and subsequently leads to the possibility of blind recovery with Algorithm 1. But first, before focusing on the case of blind recovery, the intermediate step of k-sparse recovery is introduced, that will allow more meaningful benchmarking and is simple to understand as an intermediate step towards blind recovery.

Several modifications to the general case of *dense* sensing exist, including row and column only based sensing matrices or those with a complexity of $\text{rank} = 1$

(Wagner and Zuk, 2015; Cai and Zhang, 2015; Zhong, Jain, and Dhillon, 2015). The common case of *entry* sensing, which requires additional assumptions for guaranteed recovery (Candes and Plan, 2010) and knowledge of specific entries of matrix $\mathbf{X}$, can be seen as a special case of *dense* sensing and is the simplest example of k-sparse recovery. Here, each sensing matrix is 1-sparse, e.g. contains only one nonzero entry, and the respective values are typically set to a constant. As mentioned in the overview of this approach, if sufficient quantitative standards or spike-ins are available to obtain estimates at specific nonzero entries $\mathbf{\Omega}_{(s_1,t_1)}$ of matrix $\mathbf{X}$, then *post hoc* bias recovery through *entry* sensing is possible, with $s_1 \sim \text{Uniform}(\{1,\ldots,n\})$, $t_1 \sim \text{Uniform}(\{1,\ldots,m\})$ and $\mathbf{y}_i = \mathbf{X}_{s_1 t_1}$. The necessary 1-sparse sensing matrices $\mathbf{A}_i$ are then defined as:

$$(\mathbf{A}_i)_{jk} \begin{cases} \sim 1 & \text{if } (j,k) = (s_1,t_1) \\ = 0 & \text{otherwise} \end{cases} \tag{2.1}$$

Thus, with the availability of quantitative standards defining linear operator $\mathbf{A}$ and measurements $\mathbf{y}$, the standard matrix recovery problem given in Definition 1 is then solved by Riemannian optimization (Vandereycken, 2013), specifically with conjugate gradient techniques. The resulting underling bias matrix $\mathbf{X}$ is subsequently used to normalize the database at hand.

For the case of 2-sparse recovery, a more complex example of k-sparse recovery with entries $\mathbf{\Omega}_{(s_1,t_1)(s_2,t_2)}$ chosen uniformly at random as before and $(s_1,t_1) \neq (s_2,t_2)$, sensing matrices $\mathbf{A}_i$ are then defined as:

$$(\mathbf{A}_i)_{jk} \begin{cases} \sim \mathcal{N} & \text{if } (j,k) \in \{(s_1,t_1),(s_2,t_2)\} \\ = 0 & \text{otherwise} \end{cases} \tag{2.2}$$

Analogously as in *dense* sensing and *entry* sensing described previously, k-sparse recovery presupposes a measurement setup that provides prior information about $\mathbf{A}_i$ and $\mathbf{y}_i$ to recover the underlying bias matrix $\mathbf{X}$ and is thus typically unfeasible for the normalization of high-throughput databases. It is used as an intermediate step towards the derivation of blind recovery performed with Algorithm 1. Notably, inaccuracies due to the redundancy estimation step in the blind recovery case are not an issue when benchmarking performance and robustness of the proposed algorithm under this ideal setting.

2.3.1 Blind recovery

In blind recovery it is shown how to estimate the necessary prior information from the observed signal $\mathbf{O}$ used to determine redundancies that leads to estimates for $\mathbf{A}_i$ and $\mathbf{y}_i$. With these estimates the low-rank bias matrix $\mathbf{X}$ is then recovered, which allows for the normalization of high-throughput databases. Specifically, the values for

entries $\mathbf{\Omega}_{(s_1,x)(s_2,x)}$ of 2-sparse sensing matrices $\mathbf{A}_i$ are determined by redundancy information, such as linear dependencies between features and samples estimated from a correlation matrix, which must be estimated from $\mathbf{O}$. In addition, the measurements y can be constructed indirectly from the redundancy information as outlined in Figure 2.2. The 2-sparse sensing matrices $\mathbf{A}_i$ and respective measurements $\mathbf{y}_i$ are hence defined for blind recovery as:

$$(\mathbf{A}_i)_{jk} \begin{cases} = \hat{\sigma}(\mathbf{O}_{s_1*}) & \text{if } (j,k) = (s_1, x) \\ = \hat{\sigma}(\mathbf{O}_{s_2*}) & \text{if } (j,k) = (s_2, x) \\ = 0 & \text{otherwise} \end{cases} \tag{2.3}$$

$$\mathbf{y}_i = \hat{\sigma}(\mathbf{O}_{s_2*})\mathbf{d}_{s_2} - \hat{\sigma}(\mathbf{O}_{s_1*})\mathbf{d}_{s_1} \tag{2.4}$$

where $\hat{\sigma}(\mathbf{O}_{s_1*})$ and $\hat{\sigma}(\mathbf{O}_{s_2*})$ are estimates of the standard deviation of the corresponding rows $\mathbf{O}_{s_1*}$ and $\mathbf{O}_{s_2*}$ of the observed signal, respectively. See Figure 2.1 for more details on the observed signal $\mathbf{O}$. Furthermore, $[\mathbf{d}_{s_1}, \mathbf{d}_{s_2}]$ is the orthogonal vector from point $(\mathbf{O}_{s_1x}, \mathbf{O}_{s_2x})$ to the line crossing the origin with slope $\hat{\sigma}(\mathbf{O}_{s_1*})/\hat{\sigma}(\mathbf{O}_{s_2*})$ in the space of rows $\mathbf{O}_{s_1*}$ and $\mathbf{O}_{s_2*}$. See Figure 2.2 for more details. Thus, $\mathbf{y}_i$ can be reconstructed from relative constraints encoded in the dependencies of $\mathbf{O}$. Without specifying an absolute value in the form of a quantitative standard, but by specifying a dependency, the bias can be modeled to fall on a line that runs through point $(\mathbf{O}_{s_1x}, \mathbf{O}_{s_2x})$ given that the matrix is centered (a standard assumption in matrix recovery). Since redundancies not only exist for features but also samples, the transposed observed signal $\mathbf{O}^T$ and its corresponding matrix entries $\mathbf{\Omega}^T_{(s_A,v)(s_B,v)}$ are used equivalently. Thus, while s_1/s_A and s_2/s_B specify a dependent pair of rows/columns, x/v specifies a particular observation in the space of that dependent pair of rows/columns. With linear operator $\mathbf{A}$, bias matrix $\mathbf{X}$ and measurements $\mathbf{y}$ defined accordingly in the blind recovery case, the subsequently standard matrix recovery problem akin to Definition 1 is solved by Riemannian optimization (Vandereycken, 2013) with the Pymanopt implementation (Townsend, Koep, and Weichwald, 2016). The code for the software package developed for blind recovery is provided online[2]. Specific details are given in the Supplementary Material A.

Algorithm 1 Blind Compressive Normalization (BCN)

Input: O (observed signal matrix)
Output: X (estimated bias matrix)
1: Select estimated feature and sample pairs s from estimated correlation matrix
2: Determine correlation sign d
3: Estimate standard deviations σ of each feature and sample
4: Determine measurement operator A and measurements y from s, d and σ
5: Solve for X with conjugate gradient descent based on A and y
return X

[2] http://github.com/a378ec99/bcn

2.3.2 Simulation

I this section a series of simulations are presented that evaluate the performance and robustness of k-sparse and blind recovery. To this end, a synthetic high-throughput database has been constructed by combining an underlying redundant signal with a low-rank bias to be recovered. The redundant signal $\mathbf{S}$ is generated from a matrix normal distribution, which is a common model for high-throughput data (Allen and Tibshirani, 2012).

Specifically, the signal $\mathbf{S} \sim \mathcal{MN}_{n\times p}(\mathbf{M}, \mathbf{A}\mathbf{A}^T, \mathbf{B}^T\mathbf{B})$, where $\mathbf{M}$ denotes the mean matrix and both $\mathbf{A}\mathbf{A}^T$ and $\mathbf{B}^T\mathbf{B}$ denote the covariance matrices describing redundancies in feature and sample space, respectively. Actual sampling is performed by drawing from a multivariate normal distribution $\mathbf{N} \sim \mathcal{MN}_{n\times p}(\mathbf{0}, \mathbf{I}, \mathbf{I})$ and transforming according to $\mathbf{S} = \mathbf{M} + \mathbf{ANB}$. Importantly, different features and samples have different standard deviations, which are used in combination with random correlation matrices in the construction of covariance matrices required for the construction of $\mathbf{S}$. Missing values are modeled according to missing at random (MAR) or missing not at random (MNAR) strategies. The bias to be recovered is modeled as a random low-rank matrix $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ with $\mathbf{\Sigma}$ generated from $\mathrm{diag}(\sigma_1, \ldots, \sigma_m)$. Eigenvalues are denoted as σ and sampled from a $\mathrm{Uniform}(0, 1)$ distribution (matrix rank is denoted by m). Eigenvectors $\mathbf{U}$ and $\mathbf{V}$ are obtained from Stiefel manifolds generated by the QR decomposition of a random normal matrix (Townsend, Koep, and Weichwald, 2016). Both the redundant signal $\mathbf{S}$ and low-rank bias matrix $\mathbf{X}$ are combined additively to yield the observed signal matrix $\mathbf{O} = \mathbf{X} + \mathbf{S}$. During the simulations the signal-to-noise ratio is kept approximately constant across bias matrices of different rank by scaling the eigenvalues of matrix $\mathbf{X}$ to an appropriate noise amplitude.

The performance evaluation starts with the 2-sparse sensing approach shown in figure 2.3 to 2.5 and derived in Section 2.3. This approach is closest to the general setting of *dense* sensing for which various recovery guarantees have been established (Candes and Plan, 2011). However, it differs by the random sparsification of the measurement operator (Equation 2.2). In the current setup this difference has little effect on the performance and levels off rapidly as shown in Figure 2.3. The storage requirement for the *dense* sensing approach becomes prohibitive quickly (Cai and Zhang, 2015) and therefore above 8-sparse measurement operators are not simulated. Notably, no significant difference is observed in performance between a 4-sparse and 8-sparse measurement operator in Figure 2.3.

In Figure 2.4 the advantageous scaling behavior of the 2-sparse approach is highlighted. Thus, large high-throughput databases allow reconstruction of a low-dimensional model of the underlying bias from a small fraction of potential measurements. Therefore, databases on the order of tens of thousands of features or sample require only

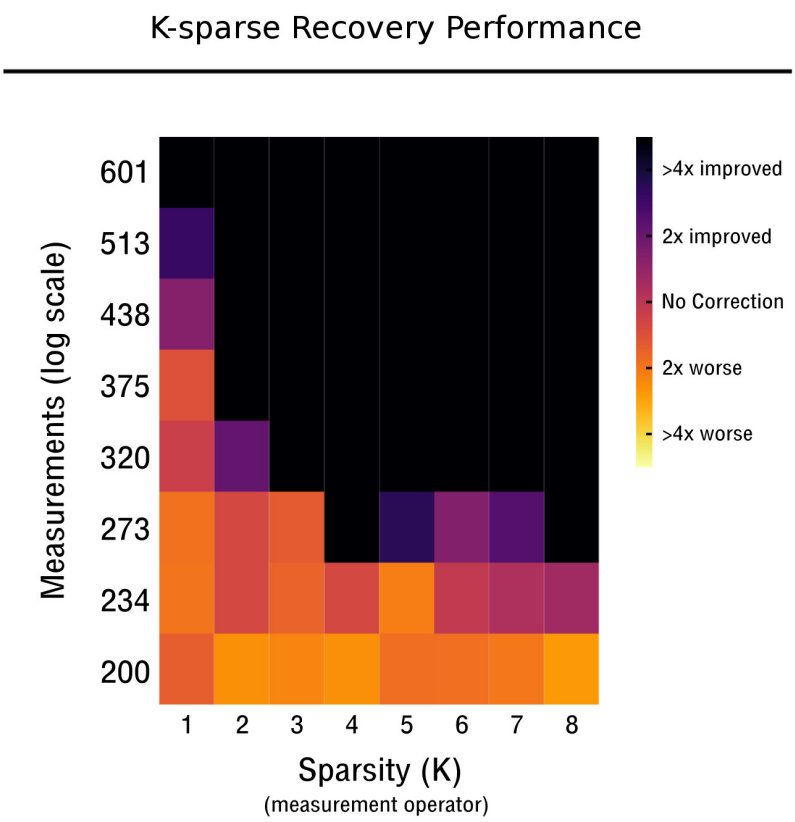


FIGURE 2.3: Performance of K-sparse recovery (Ohse, Börries, and Busch, 2019). Decreasing the sparsity of the measurement operator from 2 to 10-sparse shows a leveling-off effect in the number of measurements required for accurate recovery. High-throughput databases simulated are 50 x 50 and corrupted with a bias complexity of rank-2.

a small fraction of dependencies to be considered. When estimating dependencies for the case of blind recovery, high-specificity and low-sensitivity estimators can be used, as high-sensitivity is not necessary for an overabundance of measurements. The focus can instead be placed on high-specificity. Non-perfect recovery in the top right of Figure 2.4 is likely due to convergence failure of the conjugate gradient based solver, because of a heavily overdetermined recovery setting that is not usually applicable to the optimization routines employed. However, this can be controlled by simply dropping a fraction of the possible measurements.

In Figure 2.5 the recovery performance is shown for increasingly complex bias (rank-1 to rank-20). The necessary measurements required for improved recovery in the case of a worst-case dependency structure, e.g. max. 2500 possible measurements in Figure 2.5, are feasible to obtain up to those necessary for a noise complexity of rank-9. In the best-case scenario, e.g. max. 60000 possible measurements in Figure 2.5, measurements are feasible to obtain up to those necessary for at least

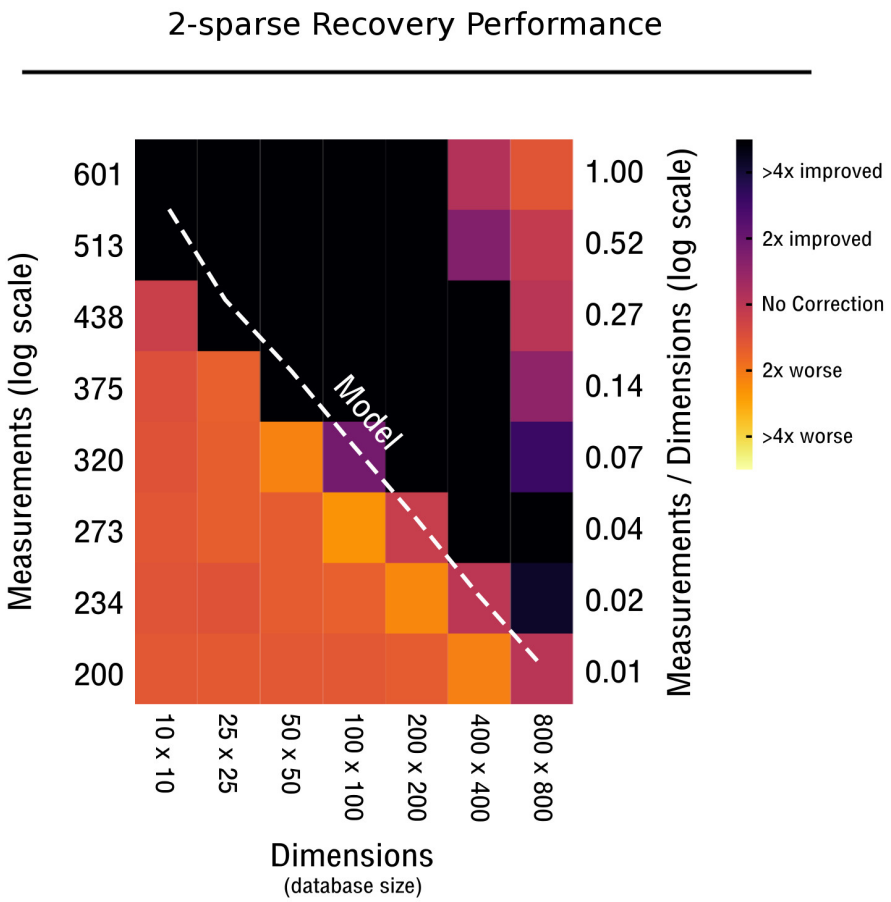


FIGURE 2.4: Performance of 2-sparse recovery (Ohse, Börries, and Busch, 2019). The scalability of 2-sparse recovery is overlaid with the theoretical model $O(c_0 r(n+m))$ (white dashed line) (Wei et al., 2016). The larger the simulated high-throughput database the more likely is reconstruction of more complex noise structures from a small percentage of measurements; the bias complexity is rank-2. Performance in the top right corner is likely decreased due to non-optimal convergence of the optimization routine in an overdetermined setting.

a noise complexity of rank-20. Notably, recovery is performed for relatively small matrix dimensions of 50 x 50 and the scaling behavior observed in Figure 2.5 may improve performance for larger setups depending specifically on the database size considered for $\mathbf{O}$.

In Figure 2.6 blind recovery performance is evaluated, where as opposed to the 2-sparse approach, entries are not sampled from a Gaussian distribution, but constructed from known or estimated dependencies. For purposes of comparisons with the 2-sparse approach, accurate estimation of dependencies is assumed. No significant difference in performance between blind and 2-sparse recovery is observed for

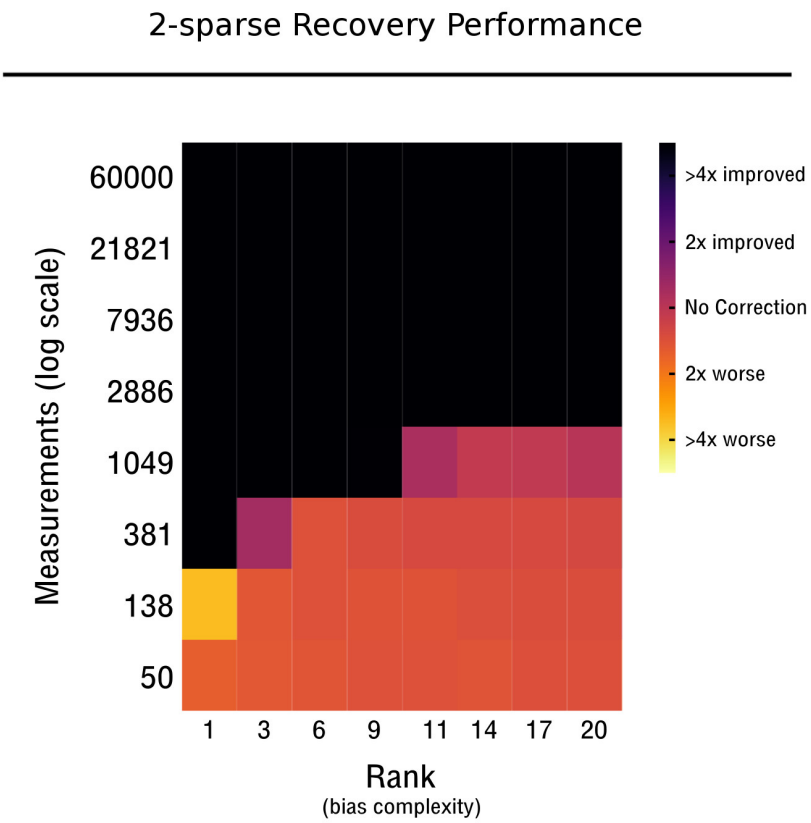


FIGURE 2.5: Performance of 2-sparse recovery (Ohse, Börries, and Busch, 2019). Proof-of-concept for 2-sparse recovery of bias with increasing noise complexity from rank-1 to rank-20 in a 50 x 50 simulated high-throughput database.

the particular setup, as shown in Figure 2.5 to Figure 2.6. Thus, recovery is feasible when the combined feature and sample space redundancies are estimated accurately and are sufficiently incoherent with the low-rank bias. Discrepancies in perfect recovery between the bottom left of Figure 2.5 and Figure 2.6 are likely due to constraints in construction of the measurement operator. Only full rows and columns are considered for blind recovery in Figure 2.6, which for matrix dimensions of 50 x 50 create measurement increments of step size 50. These do not overlap exactly with the more fine grained scale of the 2-sparse approach, leading to the observed discrepancies.

The evaluation of the blind recovery approach is extended in Figure 2.7 and Figure 2.6 with a focus on the robustness of the bias recovery. In particular, it is observed that for the case of less than ideal redundancies, such as decreased dependency strength, the bias recovery is still feasible as shown in Figure 2.7. Accordingly, as the redundant signal increases from weak dependencies ($\rho = 0.7$) to strong dependencies ($\rho = 1.0$) fewer measurements are necessary to blindly recover an unknown

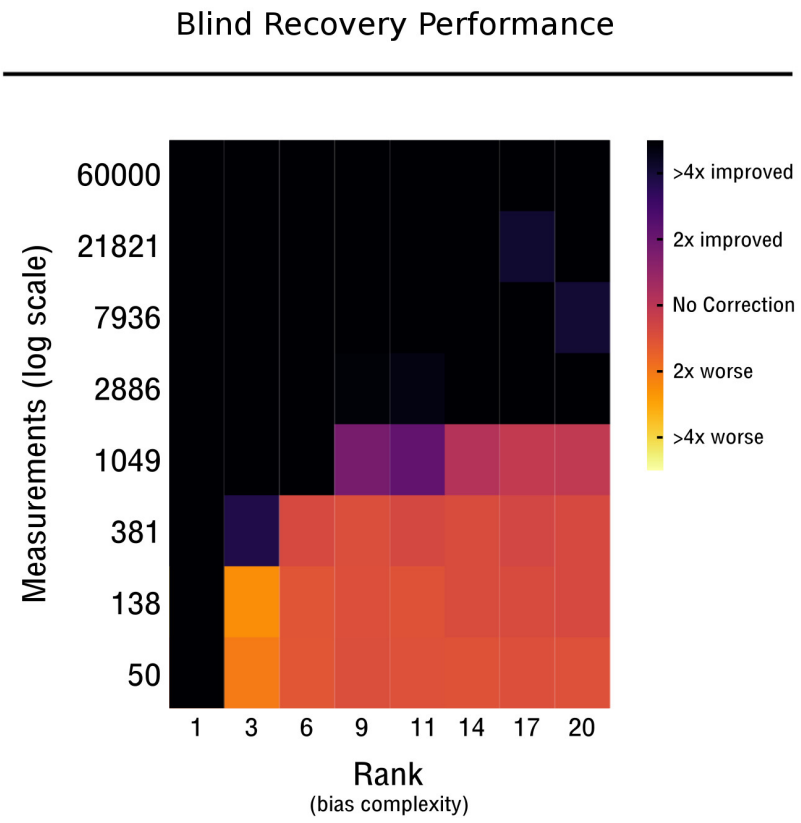


FIGURE 2.6: Performance of blind recovery (Ohse, Börries, and Busch, 2019). Proof-of-concept for blind recovery of bias with increasing noise complexity from rank-1 to rank-20 in a 50 x 50 simulated high-throughput database.

rank-2 bias for dimensions 50 x 50 in Figure 2.7. Thus, the blind recovery variant is robust to imperfect redundancies likely to be found in real world high-throughput databases.

In Figure 2.9 it is observed that lower accuracy in the form of falsely estimated redundancies (incorrect pairs of dependent features or samples) are recoverable up to a certain degree given a predetermined number of measurements. In addition, a comparison with the 2-sparse approach for a similar recovery setup is provided. Here, redundancy and estimation accuracy are mostly equivalent to additive noise in $\mathbf{y}$ (see Figure 2.8) and shuffled measurement operator A (see Figure 2.10). Both the 2-sparse measurement operator based approaches perform well in the robustness evaluation, but it is difficult to completely align the respective scales with the matching blind recovery approach, since the underlying mechanisms are different.

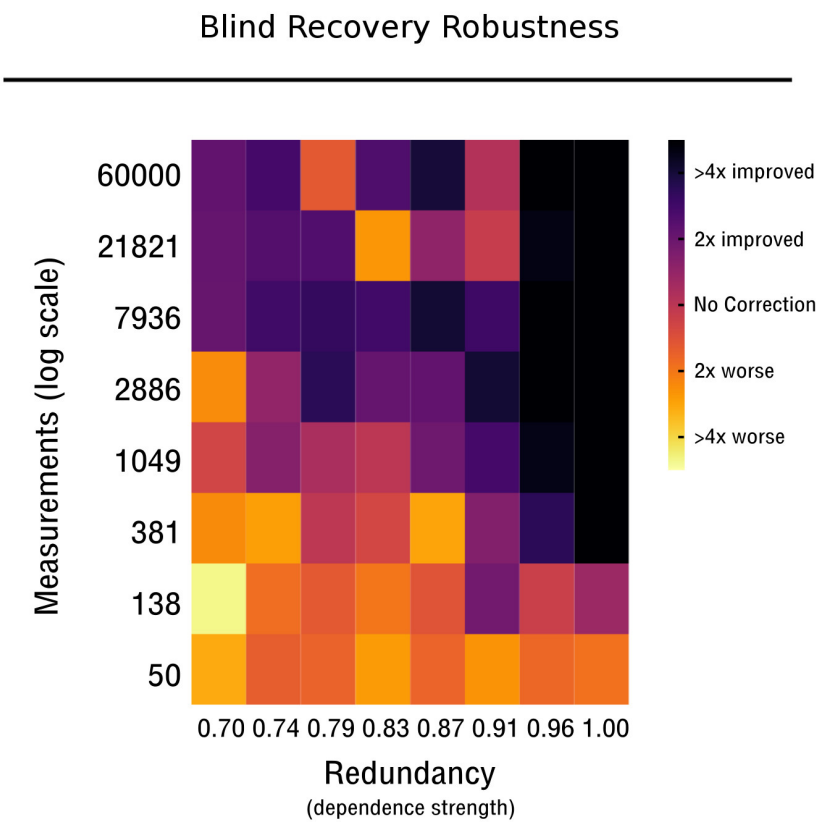


FIGURE 2.7: Robustness of blind recovery (Ohse, Börries, and Busch, 2019). As redundancy increases from weak ($\rho = 0.7$) to strong ($\rho = 1.0$) less measurements are required to blindly recover the low-rank bias. High-throughput databases simulated are 50 x 50 and corrupted with a bias complexity of rank-2. The corresponding 2-sparse recovery is simulated for additive noise in $\mathbf{y}$ or shuffling in $\mathbf{A}$ to mimic the effect of varying redundancy and estimation accuracy for the non-blind case.

2.3.3 Assumptions

For the sparsity assumption to be effective in the recovery of bias an additional two assumptions must be satisfied. First, sufficient redundancies in the form of detectable linear dependences must exist in the public database to be normalized. This assumption is generally satisfied for the complex systems measured, as the obtained RNA, protein or metabolite networks typically exhibit some strong dependences that are detectable despite the effect of confounding factors (see Chapter 1). In addition, high-throughput databases of a certain size are likely to contain redundancies in the form of similar biological samples that can be leveraged despite the effect of confounding factors. Thus, sufficient redundancies are expected in high-throughput databases.

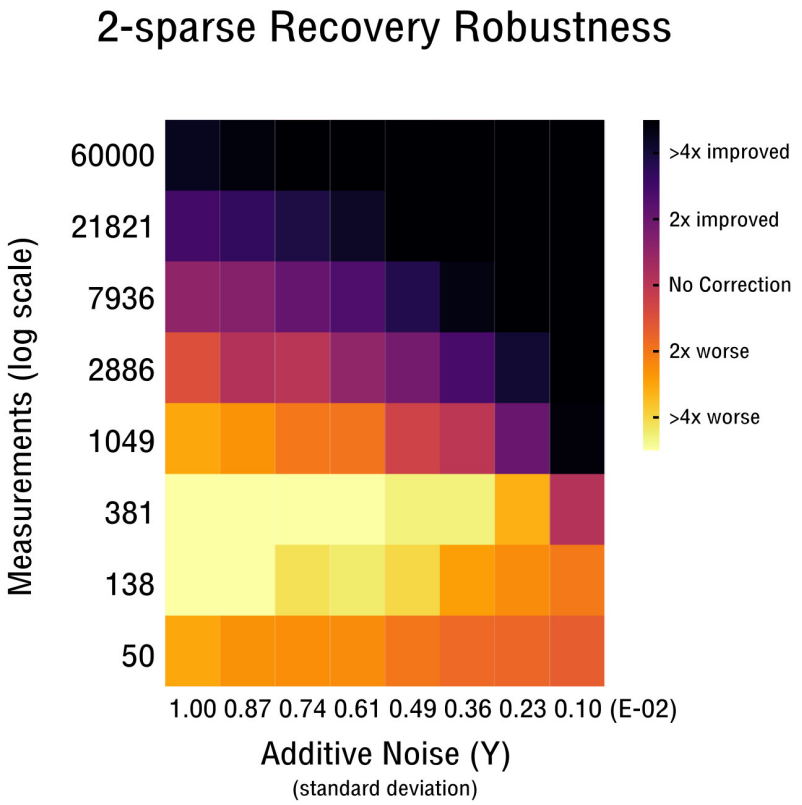


FIGURE 2.8: Robustness of 2-sparse recovery (Ohse, Börries, and Busch, 2019). As redundancy increases from weak ($\rho = 0.7$) to strong ($\rho = 1.0$) less measurements are required to blindly recover the low-rank bias. High-throughput databases simulated are 50×50 and corrupted with a bias complexity of rank-2. The corresponding 2-sparse recovery is simulated for additive noise in $\mathbf{y}$ or shuffling in $\mathbf{A}$ to mimic the effect of varying redundancy and estimation accuracy for the non-blind case.

Secondly, blind bias recovery is feasible only if the detected dependencies are sufficiently incoherent with the underlying bias modeled as a low-dimensional manifold. The likelihood of such incoherence is maximized if dependent features and samples exhibit standard deviations similar to those drawn from a normal distribution. For example, this is the case in the intermediate setting of bias recovery with k-sparse sensing matrices sampled from a Gaussian distribution. In the setting of blind recovery this assumption may only be satisfied for features and not for samples, as dependent samples generally have similar standard deviations and are unlikely to be anti-correlated in high-throughput databases. However, during the evaluation of recovery performance on high-throughput databases this does not appear to play a major role in the recovery performance (Figure 2.5, Figure 2.6 and

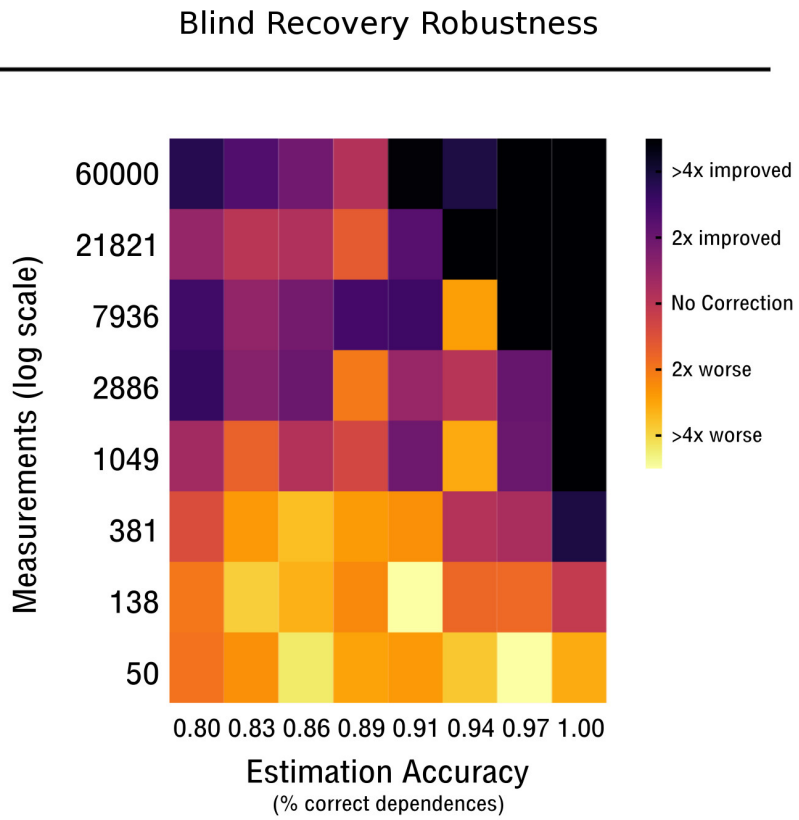


FIGURE 2.9: Robustness of blind recovery (Ohse, Börries, and Busch, 2019). As the accuracy of estimating signal redundancies from the confounded observations increases, the number of measurements required to blindly recover the low-rank bias are reduced. High-throughput databases simulated are 50 x 50 and corrupted with a bias complexity of rank-2. The corresponding 2-sparse recovery is simulated for additive noise in $\mathbf{y}$ or shuffling in $\mathbf{A}$ to mimic the effect of varying redundancy and estimation accuracy for the non-blind case.

Table 2.2). A theoretical investigation of worst case performance and recovery guarantees is still outstanding and not covered here. However, recent developments in the field of blind deconvolution and compressed sensing are in pursuit of answers to this question (Stöger, Jung, and Krahmer, 2016).

2.3.4 Optimization

The optimization routine used to solve the matrix recovery problem outlined in the preceding sections (Definition 1) is based on the theory of Riemannian manifolds. Manifolds in general are topological spaces that around each point on the manifold behave akin to Euclidean space (Munkres, 2000). This is denoted as being homeomorphic to Euclidean space. The dimensionality of a particular manifold

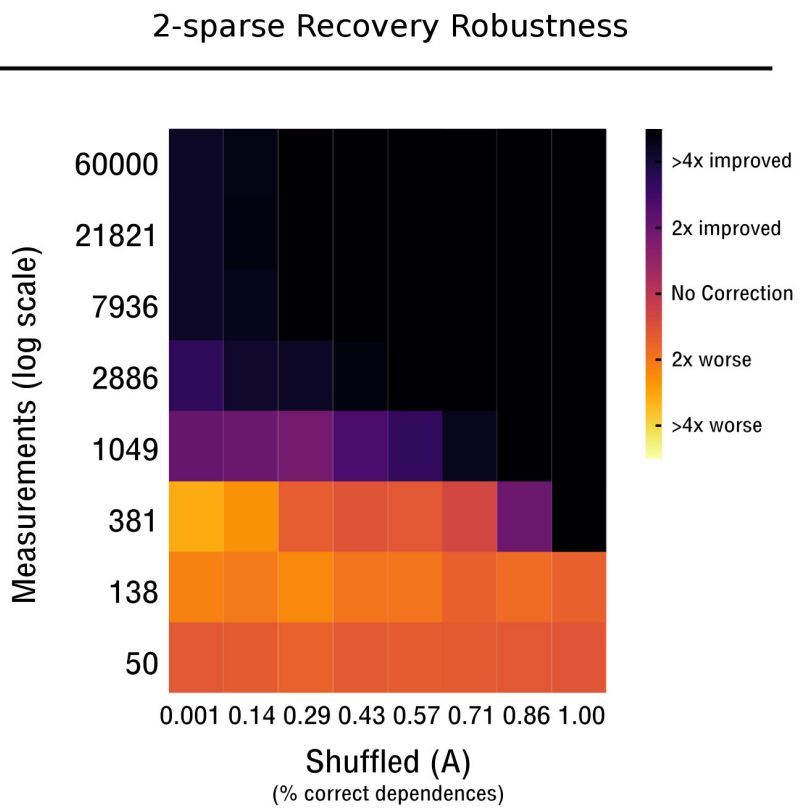


FIGURE 2.10: Robustness of 2-sparse recovery (Ohse, Börries, and Busch, 2019). As the accuracy of estimating signal redundancies from the confounded observations increases, the number of measurements required to blindly recover the low-rank bias are reduced. High-throughput databases simulated are 50 x 50 and corrupted with a bias complexity of rank-2. The corresponding 2-sparse recovery is simulated for additive noise in $\mathbf{y}$ or shuffling in $\mathbf{A}$ to mimic the effect of varying redundancy and estimation accuracy for the non-blind case.

can be much smaller than its so termed embedding dimension and includes many common shapes, such as lines, circles, spheres or more complex shapes (Munkres, 2000). Specifically, Riemannian manifolds are a subtype of manifolds that are real and smooth and exhibit an inner product on the tangent space that varies smoothly between points (Munkres, 2000).

The Riemannian optimization algorithm proposed by Vandereycken (2013) solves the matrix recovery problem in a scalable fashion. Theoretical guarantees have been developed by Wei et al. (2016). By use of the sparsity assumption the NP-hard matrix recovery problem becomes feasible for the proposed Algorithm 1. The particular advantages of Riemannian optimization are that any type of linear operator is supported and thus the recovery algorithm is not constrained to simple matrix

completion (*entry* sensing) as most current recovery algorithms are. This generalization is therefore important to solve affine matrix recovery problems more relevant to the bias recovery setting at hand. Overall, most recovery methods require a fixed rank to be known *a priory* of the low-rank matrix to be recovered. Exceptions are softImpute and softImpute-ALS (Mazumder, Hastie, and Tibshirani, 2010; Hastie et al., 2015), which iteratively increase the estimated rank until an optimal solution has been attained. Equivalently, for Riemannian optimization a pursuit version has been developed (Tan et al., 2014). This is an efficient way to deal with the unknown rank of the matrix that is to be recovered.

2.3.5 Validation

In order to validate the developed blind recovery approach we mimic a standard research problem involving high-throughput data and compare to a widely used unsupervised normalization approach. The aim is to identify differentially expressed genes under different noise conditions at a given significance level ($p = 0.05$). For this purpose a high-throughput database is simulated as defined in Section 2.3.2. It contains 30 samples with 40 measured genes (features) each and two groups of replicates that are used to determine differential expression by a standard t-test. We force accurate estimation of correlations and corresponding standard deviations, as the small database size yields poor estimates that cause the recovery to be unstable for the limited number of available measurements (see Figure 2.3, Figure 2.5). The benchmark is performed across different noise conditions: random noise derived from $\mathcal{N}(0, 1)$, systematic noise with rank-2 as outlined in Section 2.3.2 and no noise (see Table 2.1). In the case of random noise, both approaches perform similarly and are unable to reverse the effect of the corruption through normalization. Thus, no differentially expressed genes are detected at the given significance level ($p = 0.05$), which is expected. In the case of systematic noise, the blind compressive normalization (BCN) approach outperforms quantile normalization (QN) and is able to detect differential expression given the accurate estimation of correlations and corresponding standard deviations. In the case of no noise, no correction (NC) performs best, followed by the QN and BCN approach. Both approaches are able to detect differentially expressed genes for the case of no noise. Overall, this benchmark shows that the developed approach can outperform existing approaches on a standard research problem under idealized conditions.

	BCN	(avg. p-value)	QN	(avg. p-value)	NC	(avg. p-value)
Random Noise	-	3.42E-01	-	4.17E-01	-	3.89E-01
Systematic Noise	+	3.16E-02	-	1.66E-01	-	1.67E-01
No Noise	+	3.01E-03	+	3.64E-26	+	2.04E-42

TABLE 2.1: Evaluation of blind compressive normalization in an idealized setting. Comparison of blind compressive normalization (BCN) with quantile normalization (QN) and no correction (NC) of the corrupted data. Data was corrupted with random, systematic and no noise. A t-test is performed between two groups of replicates (five each) for all genes (40 in total) and the resulting p-values are averaged. Plus (+) and minus () denote if the avg. p-value falls below the significance level of 0.05, where the expected avg. p-value for no noise and no correction is 2.04E-42. A significant improvement can be noted.

To validate the developed blind compressive normalization on high-throughput data, raw transcriptome data was obtained from NCBI GEO (Barrett et al., 2013) for a range of measurement technologies, including different microarray platforms. Initial preprocessing and summarization of the obtained samples to the gene expression level was performed with SCAN.UPC (Piccolo et al., 2012). Benchmarking was performed by multi-class classification with accuracy as the metric of choice. The necessary sample labels were obtained by text mining NCBI GEO through the SQLite interface developed by Zhu et al. (2008). The dimensionality of the dataset was subsequently reduced to two dimensions with the t-SNE algorithm (see 4.2.3) in order to simplify the visualization and evaluation. Importantly, this ensures that the evaluation is not overshadowed by the strength and weaknesses of the multi-class classification algorithm used. Multi-class label predictions were performed by a naive Bayes algorithm (Pedregosa et al., 2011). Only samples that had matching annotations were used for the classification. Subsequent cross-validation was performed with a 5:1 split.

Metric	SCAN.UPC	SCAN.UPC + BCN
Accuracy (standard deviation)	0.72 (+/- 0.12)	0.82 (+/- 0.06)

TABLE 2.2: Evaluation of blind compressive normalization on high-throughput data. Standard deviations of naive Bayes based multi-class prediction accuracy overlap between the SCAN.UPC and SCAN.UPC + BCN approaches. No significant improvement can be noted.

In Table 2.2 no significant improvement of the blind compressive normalization algorithm in combination with the already quite proficient SCAN.UPC approach can be noted. This is likely due to several assumptions that are not satisfied for the limited amount of samples and features considered here (1348 samples, 84 features)

and possibly convergence failures of the conjugate gradient based solver routine. The later may be ameliorated with more restarts of the optimization routine and longer run times to reach convergence. However, if the low-rank approximation of the bias for this particular dataset is inappropriate, a larger number of samples and features with strong dependencies is required in order to be sufficient for a particular complexity of the bias to be recovered. In Figure 2.11 and Figure 2.12 the bias stemming from different measurement platforms can be seen clearly (squares and circles) and thus both normalization approaches are not sufficient at this stage. A major challenge for the large-scale evaluation of normalization algorithms is that the ground truth is generally unknown, thus it is possible that the multiple clusters of yet identically annotated cell types are accurate descriptions of the underlying biology.

As bias recovery is based on the sparsity assumption it does not necessitate the modeling of the biological signal. Subsequently, a low-rank model of the bias can be recovered, given sufficient accurate dependencies are detectable in the data at hand. The databases considered here are typically in the form of public high-throughput

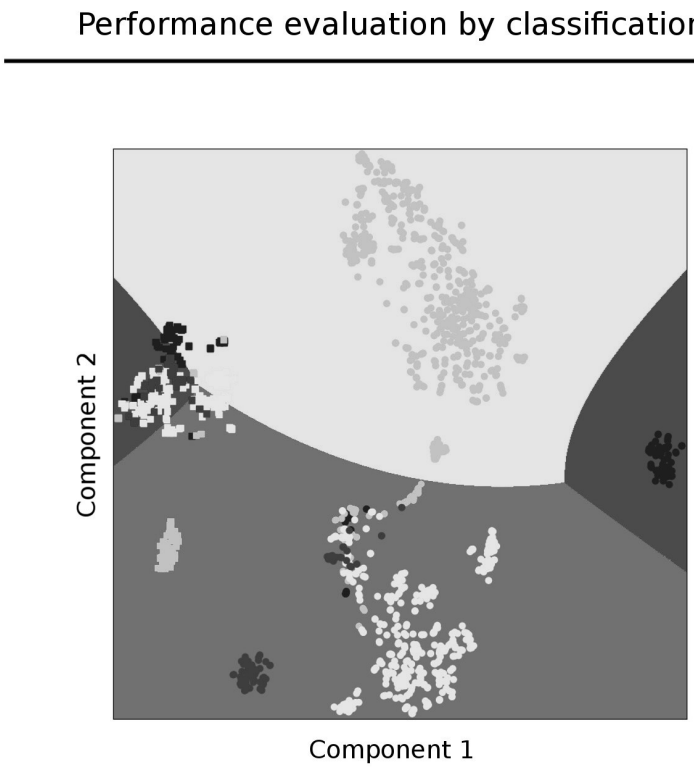


FIGURE 2.11: Decision surface of naive Bayes classification on 4 different tissue types (different shades of grey) before blind compressive normalization. Circles denote the GPL1261 microarray platform and squares the GPL570 microarray platform. A microarray platform specific bias is observable and classification accuracy is at 72% (+/- 12%).

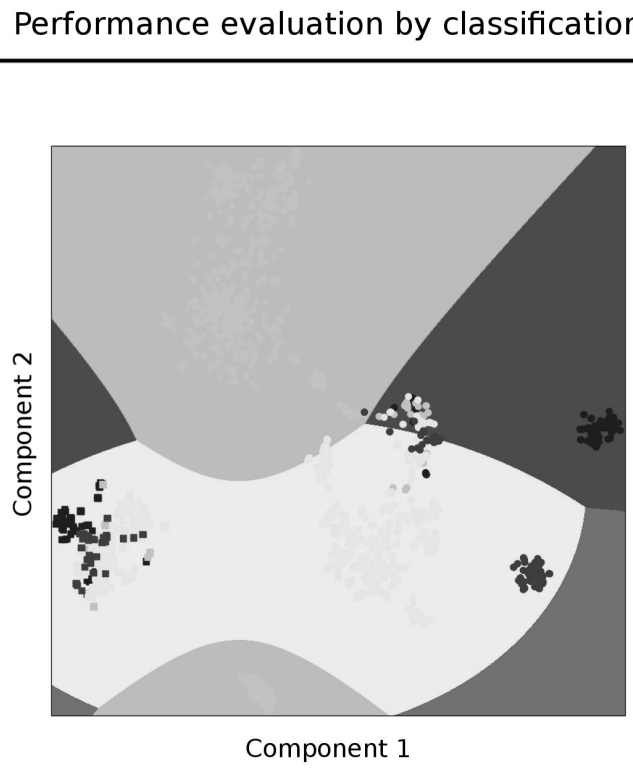


FIGURE 2.12: Decision surface of naive Bayes classification on 4 different tissue types (different shades of grey) after blind compressive normalization. Circles denote the GPL1261 microarray platform and squares the GPL570 microarray platform. A microarray platform specific bias is observable and the classification accuracy is at 82% (+/- 6%).

repositories of experiments performed in the fields of molecular and cell biology, as well as medicine. In Section 2.3.2 the proposed Algorithm 1 is validated through simulations and subsequently tested on high-throughput databases in order to determine its robustness and performance in a realistic setting (see Figure 2.3 to Figure 2.10). The lack of an effective *post hoc* bias correction technique has limited the ability to integrate public high-throughput data at scale. Such data integration is needed in order to conduct meta-analyses and improve the reproducibility and re-use of public data for individual high-throughput studies. An outlook is given in Chapter 5 with respect to future improvements and challenges of the developed algorithm in pursuit of this goal.

2.4 Conclusion

The blind compressive normalization algorithm presented in the preceding sections corrects for unknown technical and biological bias in high-throughput databases.

It does not require ad hoc assumptions about noise sources or the biological signal. Thus, the algorithm recovers systematic bias in a blind fashion by detecting and subsequently leveraging redundancies in the database at hand, such as similar samples or features. Based on these redundancies the constraints needed for bias recovery can be estimated. The algorithm addresses a need that has developed due to a lack of quantitative standards and limited reproducibility of individual high-throughput experiments. The bias recovery techniques underlying blind normalization are based on optimization routines adapted from the area of compressed sensing; specifically, Riemannian optimization routines that exploit efficient computation on low dimensional manifolds (Vandereycken, 2013). The novel contribution to the field of high-throughput data normalization outlined in the preceding sections is the framing of the problem of bias recovery as a standard compressed sensing problem and the construction of sparse measurement operators from the redundancies of a particular high-throughput database at hand. An outlook is given in Chapter 5 with respect to future work and challenges.

Chapter 3

Measure of biological relevance (MBR)

3.1 Introduction

With the rise of high-throughput technologies in the field of molecular and cell biology a large amount of high-dimensional data needs to be endowed with biological meaning. This is generally achieved through the process of hypothesis testing. But, in order to asses the biological relevance of positive or negative hypothesis tests, the domain expertise of an expert is typically required. For example, in a particular high-throughput experiment multiple genes may be significantly differentially expressed. However, only some of these may be of relevance for the biological phenomenon under study and subsequently pursued with further experiments. This determination is based on domain expertise, which is difficult to replicate in the context of reproducible scientific research and often treated as subjective and/or inferior. Therefore it has been a common occurrence that the concept of statistical significance is erroneously equated with biological relevance. While the aim is to obtain an apparently objective criteria, such an approach is inherently incorrect (EFSA, 2011). The appropriate determination of biological relevance currently still requires the judgment of an expert with domain expertise, who must consider the biological material and experimental context at hand (Lovell, 2013).

Through the use of public high-throughput databases (Figure 1.5) it becomes possible to reduce the reliance on domain experts to asses biological relevance. Much of the experience obtained when analyzing a large number of high-throughput experiments can be encoded in appropriate null distributions of the specific test statistics used. While previously an expert had to be trained by observing many experiments in a particular domain, such as in the field of cell and molecular biology, to obtain an intuition for biological relevance, such intuition can be approximated algorithmically by mining large public high-throughput databases as proposed in this chapter. The empirical null distributions of test statistics that are computed with Algorithm 2 with this goal in mind are readily transferable to the statistical analysis

of new high-throughput experiments. Importantly, p-values or multiple testing corrected p-values (q-values) obtained under inappropriate assumptions with respect to a particular null distribution can be corrected *post hoc* through the application of Algorithm 2. The corrected p-values (or q-values) that are obtained though application of the algorithm are termed r-values and are what is the main contribution here.

The subjective component of the assessment of biological relevance can hereby be reduced to a more reproducible and objective method outlined in detail in the next sections. Specifically, the large-scale testing of random experimental contrasts made possible through public high-throughput databases, allows empirical null distributions with respect to various test statistics to be computed (see Algorithm 2). With such precomputed null distributions at hand that allow the computation of the r-values, a researcher in the field of molecular and cell biology has an additional measure in his or her toolkit that indicates the biological relevance of a significant result (based on the r-values). Thereby, the approach can correct for common incorrect assumptions[1] made regarding the null distributions of individual features in high-throughput data and mimics the work of a domain expert.

In the following section the state of the field is reviewed with respect to the concept of biological relevance, specifically with respect to significance testing and measurement scales. Next, the developed algorithm and computed r-values are described in detail (see Algorithm 2). The effect of measurement scales on the proposed algorithm is investigated with respect to its effect on significance testing and the construction of appropriate null distributions. A sample of integrated features or gene sets commonly used in the analysis of high-throughput data is then annotated with estimated false positive rates (FPRs) based on the precomputed null distributions. Furthermore, validation experiments are discussed and the theoretical foundations of the developed algorithm are explored. Notably, the null distributions computed with the proposed algorithm can be readily transferred and applied to conduct significance tests on top of previously performed analyses through the use of the computed r-values. In order to provide researchers an efficient way to apply the algorithm, a software package has been developed and is provided online[2], with specific details given in the Supplementary Material A.

3.2 State of the field

Significance testing

Significance tests are aimed at quantifying the belief that a particular value of a test statistic is due to chance. According to a literature survey in the area of molecular

[1] Typically idealistic assumptions such as Gaussian null distributions.
[2] http://github.com/a378ec99/mbr

and cell biology, a common mistake in the analysis of high-throughput data is that the significance level (p-value or multiple testing corrected q-value) is equated with the probability that the alternative hypothesis is true (Lovell, 2013). This is incorrect, as the probability of observing an effect size as large or larger as the one produced by the experiment (if the null hypothesis was true) is given by the p/q-value. Thus, the p/q-value is not the probability that the alternative hypothesis is true as implied frequently in the literature (Lovell, 2013).

When hypothesis tests are conducted little or no discussion about the biological relevance of particular test statistics and their significance is generally found, even though the biological conclusions subsequently drawn in the literature are still based on an implicit definition of biological relevance. Notably, given enough measurements, most quantitative comparisons will converge to a significant result due to the increasing shrinkage of confidence intervals obtained at large sample sizes. However, even at large samples sizes which yield ample statistical power, the question of biological relevance remains to be addressed.

Another common error found in the literature is that expression levels are implicitly equated with gene or protein activity (Maier, Güell, and Serrano, 2009). Typically, gene expression profiling experiments measure the level of mRNA molecules; more mRNA transcripts generally yield more protein but not necessarily functional protein, as post-translational modifications are ubiquitous and determine the appropriate folding and function of proteins. In addition, multi-subunit complexes do not increase their activity through the up or down regulation of one particular subunit. Also, large changes in the level of molecules measured with high-throughput technologies are often equated with biologically relevant changes. But, small changes can have important effects with respect to toxicity and critical reactions in the molecular system studied. These should not be disregarded by default. Thus, the association between expression level and activity needs to be considered with care and can not be addressed merely by statistical significance testing.

A European scientific committee recommends that less emphasis should be placed on p-values and more on the experimental design and other important components of an experiment, as significance tests are only a specific component of an experiment and should not be the primary objective (EFSA, 2011). Important, too, are the definition of the nature and size of anticipated biological effects before an experiment is conducted, such that this information may be used in experimental planning. Only then can experiments be designed with sufficient statistical power to detect the anticipated effects when these occur. In general, a shift towards more reproducible reporting of experimental designs and the subsequent statistical analysis is deemed critical (EFSA, 2011).

Measurement scales

According to a literature search in the area of molecular and cell biology, few authors discuss the implications of the important issue of measurement scales. Often standard statistical techniques are used without an assessment of the appropriateness of the measurement scale used or any discussion of the interpretation of common significance tests with respect to the anticipated biological relevance in the experimental setting at hand. Throughout the literature the connection between phenotype and measurement scale is typically not considered for the optimal adjustment of the particular measurement scale employed (Stevens, 1946). For example, common benchmarks of normalization techniques are only concerned with technical artifacts and do not consider the measurement scale (MAQC Consortium, 2006). Thus, while data preprocessing and normalization may be accurate, biological interpretability suffers if the measured features can not be related quantitatively to the phenotypes at hand due to a lack of an appropriate measurement scale that corresponds with the biological relevance of a particular measured feature.

In the setting of high-throughput experiments, such as in the area of transcriptomics, proteomics or metabolomics, currently no standard scale exists which defines an appropriate measurement scale for the features at hand. Measurements are preprocessed by various algorithms (see Chapter 1) and different measurement technologies output quantitative data on different scales. Subsequently, these measurements are only comparable on the scale that a particular combination of normalization algorithm and measurement technology produces and different experiments are not inherently comparable across the same features, if the scales are not adjusted accordingly. However, even across identical measurement technologies and normalization algorithms the feature scales are not necessarily comparable when it comes to *different* features. This has lead to the widespread adoption of fold changes as primary measure of effect size, which comes with its own challenges (Lovén et al., 2012). Especially in the case of summary measures, such as those employed by gene set enrichment analyses (see Chapter 4), the integration of features and subsequent biological conclusions are highly technology and normalization dependent.

It is not clear from the literature if a feature scale can in fact be optimal for all phenotypes to be considered. This lack of clarity is inherently an issue for many high-throughput data analysis techniques that rely on the summarization of features. Different scales may be optimal for different phenotypes. Thus, even in the case where quantitative phenotypical information is widely available, the definition and implementation of a universal scale may be unfeasible. However, while there may only be measurement scales optimal on average, at minimum the reproducibility of scientific research utilizing high-throughput data can be improved by establishing such a standard scale. The focus of the developed algorithm is on null

distributions of significance tests instead of appropriate standardization of measurement scales for determining effect sizes. While these issues are inherently related, p/q-values are what is commonly reported in the literature of the field of molecular and cell biology. Thus, the focus of the developed algorithm is on determining appropriate null distributions of test statistics which can be readily applied to the reported test statistics in the literature.

3.3 Algorithm

The approach developed here can be summarized as follows. A significance test, such as a typical gene set enrichment analysis, for example with GAGE (Luo et al., 2009), calculates p or q-values and thus performs a type of t-test with or without multiple testing correction. These p or q-values are subject to the problem that they express significance but do not express the biological relevance, which is of great importance to biologists or physicians, as described in the introduction of this chapter. For example, there exist gene sets that are very often highly significant, almost regardless of the experiment that is performed and therefore actually have little biological relevance, when appearing as significant in another experiment yet again. In other words, these often significant gene sets are usually not specific to the biological phenomenon being investigated. This has different reasons, one particular being that it is commonly assumed that the null distribution is identically parameterized for each gene, which however is not the case for different genes or gene sets as observed empirically throughout the development of this approach. It is important to correct for such distinct null distributions and this is in essence what the here developed algorithm attempts. The p or q-values obtained from a particular significance test are converted through the algorithm into r-values to express a measure of biological relevance instead of significance, which is based on empirical null distributions and can be an additional measure important for researchers to consider. Specifically, for the transformation from p or q-values to r-values an empirical null distribution of the p or q-values is computed from the publicly available high-throughput data by randomized experiments and their respective gene set enrichment analyses or other significance tests commonly applied. From this empirically computed null distribution the r-values are then derived, which express a type of relative significance. As an example case, a gene set with p = 0.0001 that exhibits according to the calculated empirical null distribution in 90% of cases such a p-value or less, gets a fairly large r-value (not biologically relevant) after application of the developed algorithm. In a different case, a gene set with p = 0.01, where this p-value or less occurs only rarely in the null distribution (less than 1 % of cases), results in a fairly small r-value (biologically relevant) after application of the developed algorithm.

3.3.1 Biological relevance

The following section describes the measure of biological relevance (MBR) algorithm in detail (see Algorithm 2). It obtains an appropriate null distribution for a particular test statistic at hand and can subsequently correct misleading significance tests for a specific dataset as outlined in Figure 3.1 through the computation of r-values.

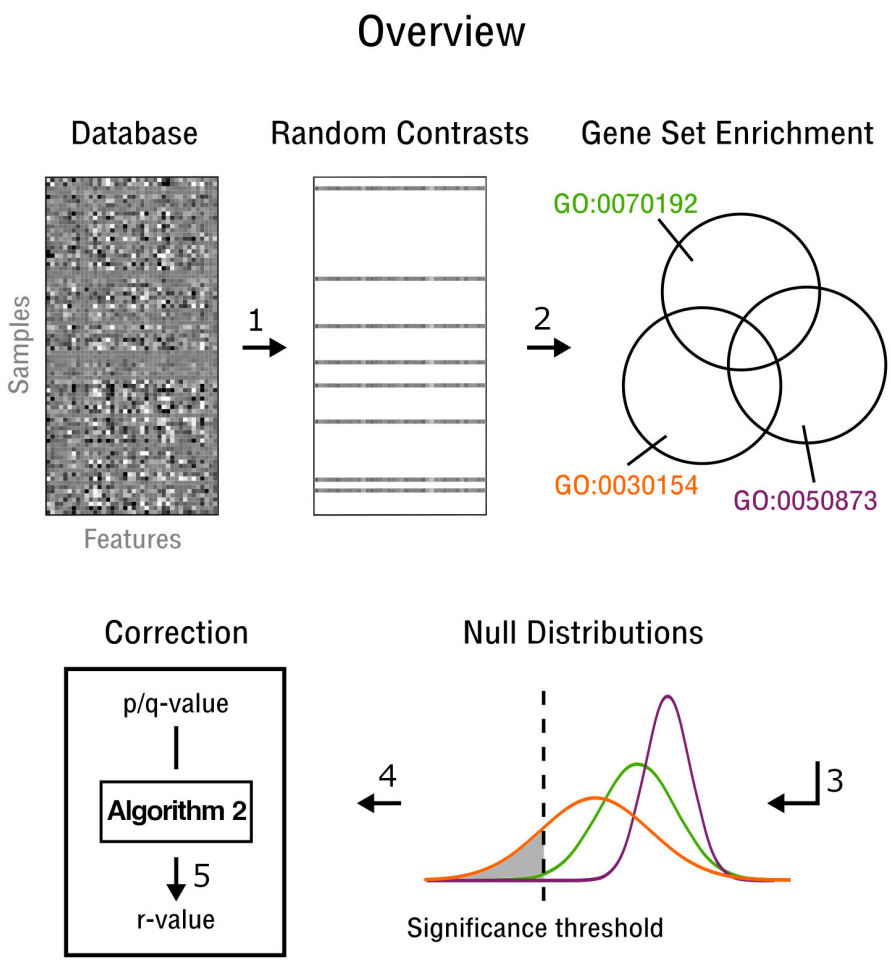


FIGURE 3.1: Measure of biological relevance (r-value) obtained through mining public high-throughput databases. Similarities are computed between samples (Step 1). Random contrasts are extracted (Step 2), which are used in combination with statistical hypothesis tests such as gene set enrichment (Step 3). The resulting empirical null distributions are fitted (Step 4) and leveraged to correct p/q-values *post hoc* for incorrect assumptions with respect to theoretical null distributions of standard significance tests (Step 5). The resulting r-values can be seen as providing information about the biological relevance of significant findings.

For this purpose random contrasts between high-throughput experiments are constructed on a massive scale, thus leveraging the information contained in public

high-throughput databases. As an important use case, the specific test statistic employed here, is a standard hypothesis test applied for gene set enrichment (see Chapter 1). Notably, gene set enrichment is common in the setting of high-throughput data analysis to describe changes in phenotypes and specifically the Gene Ontology (GO Consortium, 2004) reference is often used for this purpose. In the use case highlighted here statistical significance tests with respect to the enrichment of specific gene sets are performed with the GAGE method (Luo et al., 2009) on microarray technology (see Chapter 4). The output of the GAGE method is what is normally considered by molecular biologists in order to evaluate the biological relevance of particular gene sets (or feature combinations) when performing high-throughput analyses. The GAGE method yields p-values or multiple testing corrected q-values for each contrast. Hence the random contrasts computed in Algorithm 2 provide an empirical null distribution of these p/q-values.

Algorithm 2 Measure of Biological Relevance (MBR)

Input: p/q-value (significance measure)
Output: r-value (relevance measure)
1: Compute pairwise distances between samples in a normalized database
2: Generate random sample contrasts for particular distance range
3: Test random contrasts for significance, e.g. with GAGE (Luo et al., 2009)
4: Fit GMM to p or q-value distributions and obtain parameters
5: Use parameters to compute relevance measure of input p/q-value
return r-value

To obtain a measure of biological relevance for test statistics, such as the gene set enrichment analysis considered here, a significantly sized database of high-throughput experiments must be available in order to compute appropriate null distributions. It is important that the database used is large and diverse, to provide a realistic sample of the high-throughput experiments commonly performed. This then allows the creation of randomized contrasts between different samples and experiments.

In Algorithm 2 steps 1-4 are typically pre-computed. In step 1 of the MBR algorithm the distance between random samples a and b can be computed with the Manhattan metric (also known as the L_1 norm) given by,

$$d(a,b) = \|a-b\|_1 = \sum_i |a_i - b_i| \tag{3.1}$$

This metric is appropriate in the high-dimensional space common to transcriptomics, proteomics and metabolomics experiments.

In step 2 random contrasts are then chosen for samples within a particular distance range. Ideally, contrast are chosen from a range which does not contain samples too similar or too different, such that the null distribution is applicable to a wide

range of experiments.

In step 3 of Algorithm 2 the test statistic is computed with the GAGE algorithm (Luo et al., 2009). Multiple testing corrections yield a q-value distribution for each feature, which in the particular case described here are GO categories, but could also be genes. Selected GO categories are shown in Supplementary Figure A.4 with a focus on the diversity of null distributions that can be obtained. Table 3.1 and Table 3.2 list the most likely biologically relevant and least likely biologically relevant GO categories, respectively, obtained though Algorithm 2 (based on random contrasts). These GO categories have in addition been condensed to high level GO categories only and are subsequently shown in Figure 3.2 and Figure 3.3 to obtain a more complete overview of their relative distribution.

In step 4 of Algorithm 2 the resulting null distributions are fitted by a Gaussian mixture model (GMM) in order to facilitate easy transfer of the computed null distributions to other experiments. In this way *post hoc* significance adjustments of already obtained p/q-values can be performed without requiring a repetition of the time consuming steps 1-4 of the proposed algorithm. A GMM is defined as a probability density given by,

$$P(X|\mu,\sigma,\alpha) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(X|\mu_k,\sigma_k^2) \tag{3.2}$$

where X is the dataset of n elements $x_1,\ldots,x_n$, α_k is the mixing weight of the kth component with $\sum_{k=1}^{K}\alpha_k = 1$, $\mathcal{N}(x|\mu_k,\sigma_k)$ is the Gaussian probability density function of the kth component defined by the parameters μ_k and σ_k; with μ_k being the mean and σ_k^2 the variance of the kth component (see also Equation 3.4). The formula for an n component GMM is then given by,

$$P(X|\mu,\sigma,\alpha) = \alpha_1\mathcal{N}(X|\mu_1,\sigma_1^2) + \cdots + \alpha_n\mathcal{N}(X|\mu_n,\sigma_n^2) \tag{3.3}$$

$$\mathcal{N}(\mu,\,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{3.4}$$

where the probability density function of a Gaussian distribution (Equation 3.4) is defined by μ the mean and σ^2 the covariances.

In step 5, once the GMM fit has been obtained, these parameters can be used to determine the biological relevance of a particular q-value. Specifically, when the mixture model is viewed in the space of cumulative distribution functions instead of probability density functions, the quantile of a particular q-value can be easily obtained. These quantiles are deemed important indicators of biological relevance (r-values). For example, GO categories that are significantly enriched in a sizable

fraction of random experimental contrasts are unlikely to be biologically relevant in the experimental setting at hand. When mapping q-values to quantiles of the obtained null distributions, the value of the quantile is indicative of the ubiquity or commonness of a particular q-value for a particular GO category. For example, a q-value of 0.001 may be at a quantile of 0.2 for a particular GO category and at 00000.1 for a different GO category. Thus, the later is likely to be more biologically relevant and of interest for follow-up experiments as it is rarely obtained for random contrasts. The code for the developed software package contains pre-computed GMM parameters and is provided online[3].

Rank	FPR	GO ID	GO Description
1	42.6	GO:0006955	immune response
2	40.3	GO:0045087	innate immune response
3	39.6	GO:0051607	defense response to virus
4	38.1	GO:0006351	transcription, DNA-templated
5	37.9	GO:0055114	oxidation-reduction process
6	37.2	GO:0007186	G-protein coupled receptor signaling...
7	36.8	GO:0006355	regulation of transcription, DNA-templated
8	36.8	GO:0006954	inflammatory response
9	36.4	GO:0007067	mitotic nuclear division
10	36.4	GO:0019886	antigen processing and presentation of...
11	35.9	GO:0007067	mitotic nuclear division
12	35.7	GO:0006351	transcription, DNA-templated
13	35.7	GO:0006260	DNA replication
14	35.7	GO:0006281	DNA repair
15	35.5	GO:0006397	mRNA processing
16	35.5	GO:0008380	RNA splicing
17	35.4	GO:0008152	metabolic process
18	35.1	GO:0006099	tricarboxylic acid cycle
19	34.2	GO:0019882	antigen processing and presentation
20	34.1	GO:0006281	DNA repair

TABLE 3.1: List of gene sets frequently enriched in random contrasts. The top ranked gene set is deemed significant at the 0.05 level in 42.6% of random contrasts. This means the biological relevance of these GO categories is limited at best when determined as significant in an experiment. Random contrasts were obtained at a Manhattan distance between 8808 and 11744 and gene set sizes were on a similar order of magnitude. The top 20 gene sets listed have been observed independently in many contrasts investigated in the LungSys and GerontoSys consortia described in Chapter 4. FPR stands for false positive rate.

[3] http://github.com/a378ec99/mbr

Rank	FPR	GO ID	GO Description
-1	0.0	GO:0045368	positive regulation of interleukin-13...
-2	5.06e-03	GO:0034401	regulation of transcription by chromatin...
-3	7.58e-03	GO:0008356	asymmetric cell division
-4	1.01e-02	GO:0051594	detection of glucose
-5	1.26e-02	GO:0042420	dopamine catabolic process
-6	1.52e-02	GO:0070129	regulation of mitochondrial translation
-7	1.77e-02	GO:0051388	positive regulation of neurotrophin TRK...
-8	2.02e-02	GO:0032205	negative regulation of telomere maintenance
-9	2.28e-02	GO:0030321	transepithelial chloride transport
-10	2.53e-02	GO:0060018	astrocyte fate commitment
-11	2.78e-02	GO:0010477	response to sulfur dioxide
-12	3.03e-02	GO:0030323	respiratory tube development
-13	3.29e-02	GO:0061138	morphogenesis of a branching epithelium
-14	3.54e-02	GO:0046881	positive regulation of follicle-stimulating...
-15	3.79e-02	GO:0090188	negative regulation of pancreatic juice...
-16	4.04e-02	GO:0032369	negative regulation of lipid transport
-17	4.30e-02	GO:2001180	negative regulation of interleukin-10...
-18	4.55e-02	GO:0016344	meiotic chromosome movement towards...
-19	4.80e-02	GO:0022612	gland morphogenesis
-20	5.06e-02	GO:0042063	gliogenesis

TABLE 3.2: List of gene sets rarely enriched in random contrasts. The top ranked gene set (-1) is deemed significant at the 0.05 level in none of the random contrasts. This means the biological relevance of these GO categories is very high when determined as significant in an experiment. Random contrasts were obtained at a Manhattan distance between 8808 to 11744 and gene set sizes were on a similar order of magnitude. FPR stands for false positive rate.

3.3.2 Assumptions

High dimensional data of complex systems generally exhibits various dependencies (see Chapter 1). Biological systems such as the cell specifically manifest a high degree of organization through complex dependency networks of genes, proteins and other molecules. Therefore the independence assumptions often made for standard hypothesis tests does not hold, as these are not developed for such high-dimensional data. In addition, because some feature combinations naturally vary more strongly between or within experiments than other feature combinations, the identical null distributions of the test statistics commonly employed in gene set analysis do not provide an accurate perspective on the biological relevance. Thus, a feature specific null distribution is created with the proposed algorithm for a particular test statistic at hand. A completely randomized database where features are randomized as well as samples, for example, may not be appropriate either, as future experiments are unlikely to be drawn from such a distribution. More precisely, many distributions can be employed as null distributions. But, in order to obtain a measure of biological relevance, only cell states which exist in the real world must be used. Therefore, it is important to use actual measurements of cell states as proposed here.

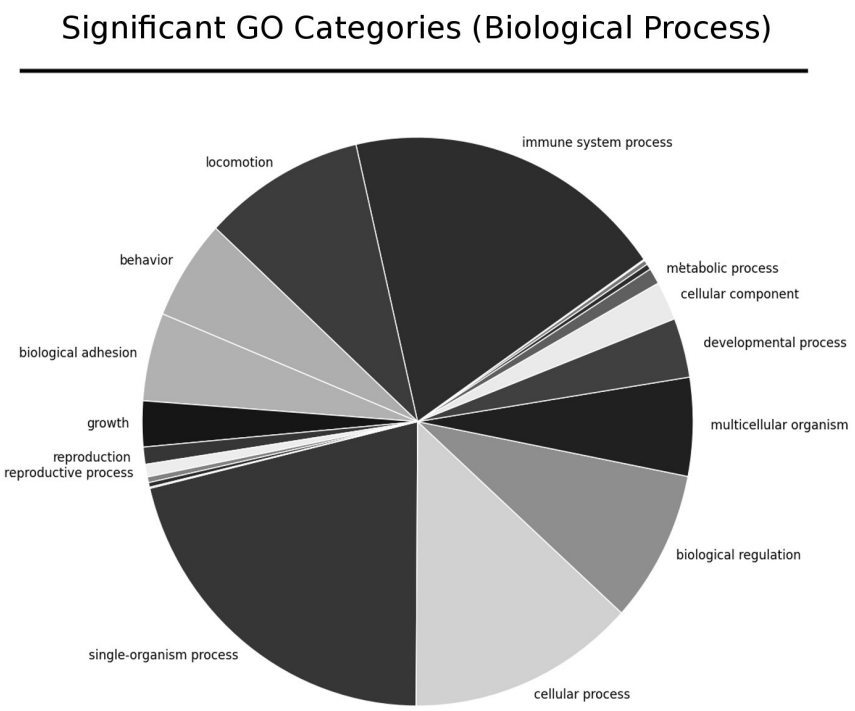


FIGURE 3.2: Proportions of significant high-level GO categories with respect to those significantly enriched across random contrasts in the GPL1261 high-throughput platform Barrett et al., 2013. The focus is on the biological process subgroup of GO categories of which certain categories dominate with respect to the proportion of significantly enriched gene sets

A major assumption must be satisfied for the proposed measure of biological relevance to yield an informative metric based on the available high-throughput data. Public high-throughput databases used to determined an appropriate null distribution of a specific test statistic must be a sample from the global distribution of experiments where future test statistics are obtained from. These databases might be oversampled for particular cell states with respect to some true underlying distribution, but such a distribution is generally unknown or unfeasible to obtain. Thus, with the assumption that future test statistics are conducted on samples drawn from a particular null distribution, cases of oversampling are secondary. In other words, it can be assumed that future experiments performed in the field of molecular and cell biology will be drawn from a similarly (potentially oversampled) distribution as past experiments. Overall, it is challenging to prove an optimal null distribution, because this may depend on the particular context and research question of interest.

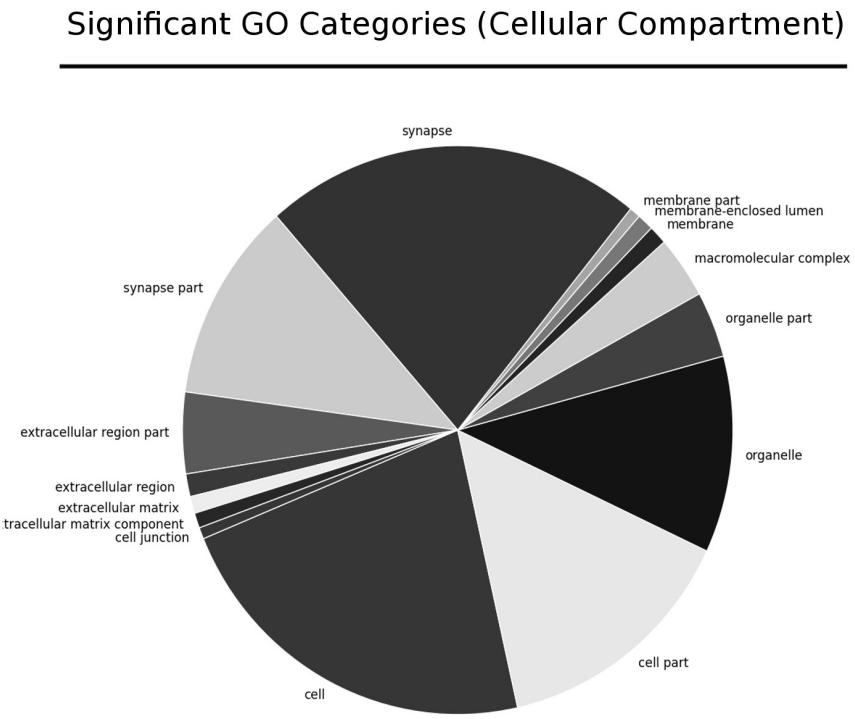


FIGURE 3.3: Proportions of significant high-level GO categories with respect to those significantly enriched across random contrasts in the GPL1261 high-throughput platform Barrett et al., 2013. The focus is on the cellular compartment subgroup of GO categories of which certain categories dominate with respect to the proportion of significantly enriched gene sets

3.3.3 Measurement scale

With the main focus on hypothesis testing and statistical significance, the matter of effect sizes and appropriate measurement scales is often neglected in the field of molecular and cell biology (see Section 3.2). However, the relation between a phenotypical effect and the scale on which quantitative measurements are obtained is critical for high-throughput experiments. Summary features are commonly constructed for an analysis, as is the case in the workflows outlined in Chapter 4. In this setting, a large number of quantitative features needs to be integrated to study the underlying biological mechanisms. Thus, the mapping from measurements of many different features to a categorical phenotype is a common task. For this integration an optimal scale of each feature measured becomes an important concern. In the case of gene set enrichment analyses, where a combination of multiple quantitative features is the goal (see Section 4.2.3). Subsequent biological conclusions are often highly dependent on the individual feature scales used. One approach to deduce an optimal scale of a measurement is to have a quantitative description of the phenotype of interest at hand. Then, a clear mapping between the considered phenotype

and a particular feature can be obtained through optimization of the measurement scale.

Several common limitations of research in quantitative biology are holding back such an approach. First, quantitative descriptions of phenotypes are rare. Most phenotypes are assessed categorically and thus it is often hard to define what an optimal scale is. Furthermore, given several quantitative descriptions of the phenotypes a particular system under study can exhibit, the question of how to combine these quantitative phenotypes or how to choose the most important phenotype for the experiment at hand becomes difficult to address. For example, a common task is the measurement of multiple genes simultaneously in a transcriptomics experiment. The subsequent association to the observed phenotypes, such as migration and differentiation, must then be performed. Because there exist generally multiple phenotypes of which it is not clear *a priori* which is of interest, the definition of a measurement scale becomes particularly difficult. Lastly, the specific features of the measured high-throughput data to be used are unclear, or specifically how to combine these appropriately into a summary measure. Therefore, the developed Algorithm 2 is based on null distributions of p/q-values and not effect sizes. However, how features are combined still plays an important role in the assessment of statistical significance, as different combination strategies lead to different hypothesis test outcomes.

3.3.4 Optimization

Fitting the obtained null distributions is performed with an Expectation-Maximization approach. This can be applied to fit any particular mixture model, but the focus here is on Gaussian mixture models described in Section 3.3.1 and denoted by Equation 3.4. Expectation-Maximization is a natural generalization of a maximum likelihood estimation strategy to the case where data is not labeled with respect to the mixture components of origin (Do and Batzoglou, 2008). An Expectation-Maximization approach consists of 3 main steps:

1. Initialization
2. Expectation (E-step)
3. Maximization (M-step)

Step 1 is performed by setting the parameters defined in Equation 3.5, Equation 3.6 and Equation 3.7 to initial estimates. In order to start from an effective initial estimate, often K-means clustering (see Section 4.2.3) is performed.

$$\mu_k = \frac{\sum_i^{N_k} x_{i,k}}{N_k} \tag{3.5}$$

$$\sigma_k^2 = \frac{\sum_i^{N_k}(x_{i,k} - \mu_k)^2}{N_k} \tag{3.6}$$

$$\alpha_k = \frac{N_k}{N} \tag{3.7}$$

with parameter α_k, where k denotes the mixture component, μ_k being the mean, σ_k^2 the variance and N_k the number of data points of the kth component. N denotes the total number of samples and $x_{i,k}$ a particular sample of component k. Step 2 and step 3 subsequently alternate until convergence of the approach. The cost function $P(X|\mu, \sigma, \alpha)$ is defined in Equation 3.8, which measures the fit of the GMM for the current parameter estimates and determines when convergence has been reached.

$$P(X|\mu, \sigma, \alpha) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(X|\mu_k, \sigma_k^2) \tag{3.8}$$

In the E-step the expectations are updated according to Bayes' rule (Equation 3.9), by estimation of the posterior probability $P(x_i \in k_j|x_i)$ according to Equation 3.11, Equation 3.11 and Equation 3.12 plugged into Bayes' rule.

$$P(x_i \in k_j|x_i) = \frac{P(x_i|x_i \in k_j)P(k_j)}{P(x_i)} \tag{3.9}$$

$$P(x_i|x_i \in k_j) = \mathcal{N}(x_i|\mu_{k_j}, \sigma_{k_j}^2) \tag{3.10}$$

$$P(k_j) = \alpha_{k_j} \tag{3.11}$$

$$P(x_i) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(x_i|\mu_k, \sigma_k^2) \tag{3.12}$$

In the M-step Equation 3.13, Equation 3.14 and Equation 3.15 are updated consecutively. These were initially estimated in step 1, but are now iteratively updated based on the preceding E-step. Lastly, convergence is reached when the most recent iteration of the approach does not lead to changes larger than a specific threshold parameter. Then, parameters for the mixture components can be extracted and used for subsequent modeling of the null distributions of p/q-values as proposed in the preceding sections.

$$\mu_k = \frac{\sum_i^N P(x_i \in k_j|x_i)x_i}{\sum_i^N P(x_i \in k_j|x_i)} \tag{3.13}$$

$$\sigma_k^2 = \frac{\sum_i^N P(x_i \in k_j|x_i)(x_i - \mu_k)^2}{\sum_i^N P(x_i \in k_j|x_i)} \tag{3.14}$$

$$\alpha_k = \frac{\sum_i^N P(x_i \in k_j|x_i)}{N} \tag{3.15}$$

3.3.5 Validation

How can a measurement scale be validated to perform superior to another measurement scale? In the approach developed here, quantitative gene sets are constructed from various features through simple addition of measured values, as is common in the case of high-throughput data. Depending on the measurement scales used for the various features, different summary values of the combined features are subsequently obtained. The combined features, in this case GO categories (see Section 4.2.3), are then categorized as active (above average) or inactive (below average) based on the combined average for all samples found in the public high-throughput database at hand.

Thus, different measurement scales result in distinct activation profiles of the constructed gene sets evaluated in Figure 3.4. Subsequently, a portion of samples from the high-throughput database was annotated with a publication link via PubMed (Doms and Schroeder, 2005). These publications were then text-mined for the occurrence of GO categories. This allowed a characterization of potentially biologically relevant GO categories for the available publications and was compared to the gene sets identified as active based on the high-throughput database at hand.

With this strategy receiver operator curves were obtained to contrast the two distinct scales used. The first scale is the commonly used log-transform based scale, while the second scale is a rank based scale, leading to values uniformly distributed on the range of 0 to 1. The rank scale is based on all samples contained in the particular high-throughput database used. The database used was derived from NCBI GEO (Barrett et al., 2013) and preprocessed according to Section 4.2.1. In this case the GPL1261 platform was the primary source of raw high-throughput data.

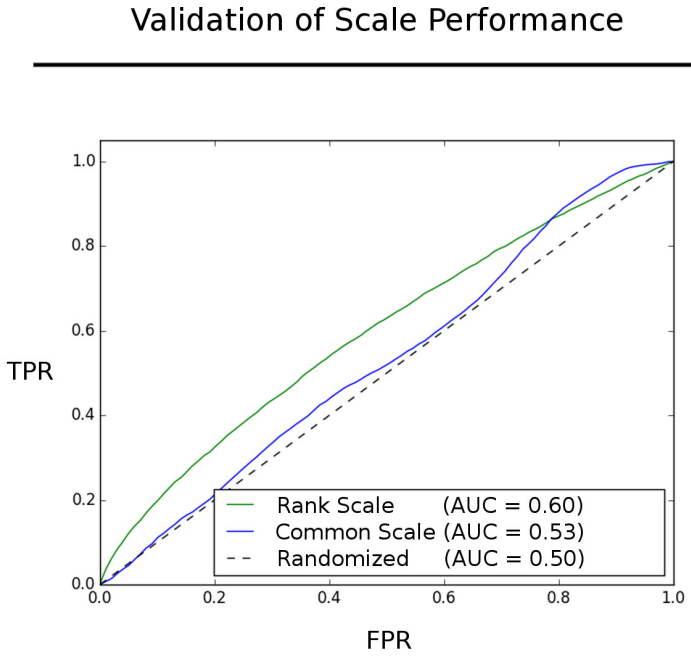


FIGURE 3.4: Receiver operating characteristic curves of a rank based scale versus the commonly used log-transform based scale. Important GO categories are predicted and validated via text-mining (gold standard). Only GO categories above the estimated noise threshold are used. The rank based scale slightly outperforms the common scale.

Notably, for any validation with respect to an optimal feature scale, it must also be assumed that the obtained sample of integrated features used (in this case GO categories) is from an appropriate null distribution that is not biased. For example, if the described text mining of publications was biased towards particular integrated features, the evaluation thereof would also be biased towards the performance on these feature combinations or GO categories (see Section 4.2.3). It was not possible to exclude this possibility.

3.4 Conclusion

The algorithm proposed in this chapter determines a measure of biological relevance that aims to provide meaningful information to researchers in the field of molecular and cell biology in addition to the typical significance measures. The main challenge addressed here is the identification of biologically relevant experimental findings based on an evaluation of the scores obtained by standard hypothesis testing. No objective metric to gauge the biological relevance of test scores currently exist outside of the advice of domain experts. However, information regarding biological relevance is critical to the study of biological mechanisms. By mining public

high-throughput databases and computing an empirical null distribution of p or q-values for common test statistics, an objective metric is obtained to gauge biological relevance, defined as the r-value. With the developed algorithm information contained in public high-throughput databases is thus leveraged akin of transfer learning. Given a particular p or q-value a *post hoc* correction can be applied that controls for inappropriate independence assumptions with respect to the null distributions. The algorithm is validated through simulation and tested on high-throughput experiments in order to determine its effectiveness in a realistic setting. The contributions outlined in the preceding sections contribute to the advance of high-throughput data integration. An outlook is given in Chapter 5 with respect to future work and challenges.

Chapter 4

Applications

4.1 Introduction

This chapter contains a description of the experiments performed within the framework of the GerontoSys[1] and LungSys[2] consortia (see Section 4.3.1 and Section 4.3.4) in addition the workflows developed as part of the conducted bioinformatic analyses. The focus of both consortia is the study of mammalian cells under multi-factorial perturbations and the integration of quantitative measurements obtained as a result of high-throughput measurements. Such a systems biological approach (see Chapter 1) and interdisciplinary perspective on cellular changes was successful at determining a novel biomarker for a specific subtype of lung cancer (Ohse, 2018) and in proposing several mechanisms that underly aging related disease (Kalfalah et al., 2015; Kalfalah et al., 2014).

The workflows developed for the two consortia and subsequent in-depth bioinformatic analyses described in this chapter, were able to achieve their impact by translating high-dimensional quantitative measurements into actionable biological insights. These insights were subsequently validated *in vitro* and *in vivo* through experiments exploring the underlying biological mechanisms in further detail. Important shortcomings with respect to current high-throughput data analysis methods were noticed throughout this process and have lead to the conception and refinement of the developed algorithms outlined previously in Chapter 2 and Chapter 3.

While the GerontoSys consortium studies the effect of aging in skin cells and its potential underlying mechanisms and the LungSys consortium studies the effect of growth factor stimulation on lung cancer cells and their observed resistance to therapy, the workflows developed and applied throughout both consortia are consistent and independent from the specific biological questions addressed. This is an advantage with respect to the reproducibility of the here conducted research. The developed workflows are outlined in the next section. Subsequently, the research

[1] BMBF #0315576D
[2] BMBF #0316042G

findings of each consortium are discussed within specific case studies (see Section 4.3.2 and Section 4.3.4).

4.2 Workflows

For the mentioned consortia several workflows concerning the analysis of high-throughput measurements have been developed. These are summarized in this section. Apart from the initial preprocessing and normalization routines, the workflows can be applied independently across transcriptomics, proteomics and metabolomics measurements. Thus, once measurements of a particular sample have been converted into the correct format and have been preprocessed appropriately, the subsequent workflow steps are identical. Since the main focus of the two consortia lies in the analysis of transcriptomics measurements, this technology is also a focus of this section.

From a bioinformatic perspective the complexity of implemention of high-throughput workflows can vary. But most importantly, a specific data analysis workflow must be able to deal with missing values and otherwise non-standard experimental designs, which often include insufficient replicates or controls and incoherent annotation. To control for the later, a common sense check needs to be performed, by which the resulting biological finding or interpretation thereof can be confirmed as reasonable. This generally requires a major amount of experience and domain expertise of the bioinformatician and is typically an iterative process.

In order to understand and characterize cellular changes within the scientific terminology of molecular and cell biology, a mapping from quantitative high-dimensional measurements to qualitative and categorical phenotypes must be performed. Thus, a major challenge in the development of workflows for the analysis of high-throughput data is the eventual summarization of the acquired measurements into information that can be understood and processed by domain experts. For this purpose, a mapping between quantitative values of thousands of features (e.g. genes or proteins) to categorical labels in the form of phenotypes has to be performed. This mapping is challenging for various reasons (see Section 4.2.3). However, since scientific exchange and communication generally takes place on the categorical level, such a mapping from high-throughput measurements to phenotypes is essential and the goal of any workflow outlined in this chapter.

4.2.1 Preprocessing

A typical workflow is to start with the preprocessing of raw measurements obtained from a high-throughput measurement instrument. This can include next-generation sequencing technology, as well as microarray or mass spectrometry instruments. Overall, multiple levels of preprocessing are usually performed.

Instrument

The first level of preprocessing is conducted typically during the measurement acquisition itself, especially in the case of transcriptomics experiments with measurements obtained by microarray technology. The exposure time of the imaging device is adjusted according to the fluorescence level observed over a particular set of samples (some microarrays contain multiple samples). Then, the measured signal intensities are binned into a mask, where the intensity in the center is assumed to be the highest peak intensity and taken as final measurement by the instrument software (Affymetrix, 2007). Subsequently, the next level of preprocessing involves mean centering of the various bins randomly selected across a microarray. Since measurements of individual mRNA fragments are assumed to be distributed randomly on the chip these should not differ over a large enough area in average intensity. Next generation sequencing, as well as proteomics approaches require different strategies to map from the detected sequence reads or peptide fragments to genes and proteins, respectively.

Summarization

For summarization of probes or sequence fragments into genes, multiple intensity values are reduced into a single value by averaging or weighted averaging (McCall, Bolstad, and Irizarry, 2010). At this stage, probes or sequence fragments which do not meet quality standards, for example due to sequence mismatches or other corruptions, are generally eliminated (Sandberg and Larsson, 2007).

Imputation

Missing value imputation is an important task in workflows that analyze high-throughput experiments. Often, a small fraction of the measured features are faulty and therefore missing in the dataset, frequently due to detection limits of the instrument or other inconsistencies in sample processing. For techniques that are only designed to work with complete measurements, missing values are a significant challenge. This typically means that the data is not usable in the form it has been obtained. Therefore, missing value imputation techniques have been developed. An important consideration for such imputation techniques, that improves reconstruction of missing values profoundly, is the modeling of the underlying factors that cause or contribute to missing values (especially how these are related to other factors considered in the experimental design). Three underling classes of missing values exist: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). However, the case of MCAR is an idealistic class that does not exist much outside of statistical theory. Experimental data that has approximately MAR values is assumed to be caused by specific factors that are yet unknown to the experimenter. If a dataset has values distributed MCAR, the missing values are biased due to censoring or other causes. This is the most difficult type of missing data

to address with imputation, as it often requires appropriate modeling of the underlying bias, which is in this case unknown. Model based imputation techniques can include many specialized approaches. A recently developed technique based on singular value decomposition (SVD) that scales to large data matrices that as are often found when analyzing high-throughput data is softImpute (Mazumder, Hastie, and Tibshirani, 2010). This technique leverages a sparsity assumption and convex relaxation techniques to reconstruct a low-rank model of the underlying signal, which is then used to impute the missing values. More simple statistical approaches include median or mean based imputation or setting all missing entries to a constant value. However, the later method can introduce significant bias in the data.

Annotation

Often times the annotation of high-throughput measurements to be analyzed is severely inconsistent. There are frequently cases where the annotation is incorrect, non-existent or partially missing. Thus, it is important to apply a certain degree of cross referencing before starting any analysis workflow. This can be performed by determining true positive and false positive controls in the dataset at hand based on prior information. For example, due to the biological background information that a certain gene or protein should be up-regulated, this can be verified in the data. Alternatively, in specific cases where samples are derived from male and female sexes, genes or proteins expressed on the Y chromosome are only to be found in the male specific samples. When it comes to larger high-throughput databases, another challenge is the processing of incoherent annotations. Here, many times multiple labels exist for the same phenotype and spelling mistakes or abbreviations make machine processing difficult. Thus, a natural language processing (NLP) approach needs to be applied that allows for regular expressions and dictionaries of biomedical terms to be incorporated in the process. Also, the to be annotated high-throughput data is typically not accessible in a comprehensive form, except for the NCBI database GEO, where an SQL based tool has been developed (Zhu et al., 2008), making machine learning based approaches challenging.

4.2.2 Normalization

Normalization begins in conjunction with the preprocessing of the raw high-throughput measurements obtained. Several approaches have been designed for this step, ranging from unsupervised to supervised machine learning based techniques. The unsupervised approaches that have been developed typically make use of ad hoc assumptions about noise sources or biological signal. These are then leveraged in an attempt to average out biological or technical bias. However, unsupervised approaches generally fail to exploit additional or prior information and it is difficult

to assess the validity of most underlying assumptions. Nonetheless, common normalization approaches are based on unsupervised learning. A frequently used approach is quantile normalization, which is part of the RMA method (Irizarry et al., 2003). The underlying assumption of quantile normalization is that the overall biological signal does not vary significantly across samples and therefore equivalent scaling across samples is deemed beneficial (Bolstad et al., 2003). This method is currently used for normalizing transcriptome and proteome measurements, including in the two case studies described in Section 4.3.2 and Section 4.3.4. However, the developed blind compressive normalization algorithm of described in Chapter 2 aims to provide an additional unsupervised approach to the normalization of high-throughput measurements that does not require the assumption that the biological signal does not vary significantly across the considered samples.

4.2.3 Analysis

Gene set enrichment

One major approach to summarize large amounts of quantitative information into categorical information (phenotypes) is to use ontologies, such as the Gene Ontology (GO) in combination with enrichment testing. The underlying assumption is that there exists a background distribution of GO categories (Ashburner et al., 2000; Consortium, 2015; Subramanian et al., 2005) annotated to a particular high-throughput dataset, against which a statistical significance test can be performed. For example, a transcriptomics experiment measures a multitude of genes, which are annotated with corresponding GO categories (in most cases multiple per gene). Then, an enrichment test is performed to see if genes from the experiment which are of interest have a disproportionate amount of a particular subset of the GO categories used. This allows the reduction of quantitative measurements to a subset of categorical GO categories. These GO categories generally describe the phenotype of the cellular system under study and are informative for biological researchers.

Ontologies

To standardize phenotypes semantically, different ontologies have been created by expert curators or in some cases though text mining of the scientific literature. The most used ontologies are GO and Reactome (Croft et al., 2014). But, there are also organism or disease specific ontologies that in specific cases are preferred. The pathways or phenotypes given are then mapped to a set of genes that is associated with these categories. This is referred to as a gene set. There exist specialized subsets of ontologies that focus on molecular function, biological processes and cellular compartments associated categories. However, the associations are tenuous at best, because there is often no clear functional relationship. Also, ontologies are not generally hierarchical, which makes subsequent analyses challenging. This is because if

one is interested in more or less specific phenotypes, it is difficult to map up or down the ontology graph without a clear hierarchy. However, some approaches such as GO slim (GO Consortium, 2004) exist to address this challenge, but are not well established yet.

Hyper-geometric testing

In hyper-geometric testing a set of genes is identified to be of interest. For example, as significantly differentially expressed. The background distribution of all remaining genes is tested against and the respective GO categories are identified as before. Thus, the set of genes that is of interest is tested via enrichment analysis.

GAGE

A more refined method is the Generally Applicable Gene set Enrichment (GAGE) approach (Luo et al., 2009), which can also detect small systematic changes that are missed by standard enrichment tests. This is often the case as global cellular changes are often subtle in nature. For example, GO categories of genes that might not be differentially expressed, but are varying slightly in the same direction in unison can be detected with this method.

Differential expression

In differential expression analysis typically a t-test (Student or Welch) is performed to asses if a gene is significantly differentially expressed. A commonly used software package from the R programming language is the LIMMA package (Ritchie et al., 2015) that conducts such statistical tests. It employs the empirical Bayes procedure, which leverages estimates of the overall distribution of the analyzed dataset and then uses this distribution as a prior for the estimation of each individual hypothesis, such as during the test for differential expression of each individual gene.

Multiple testing

When performing hypothesis tests in the setting of high-dimensional data with a large number of features, such as genes, proteins, or metabolites, the probability that a significance test produces a positive result by chance for one of the features under consideration is large. This is due to the fact that while the probability of a false positive result should be no more than 5% for a p-value of 0.05, if performed for a large number of features, the absolute number of false positives tests at the 5% threshold is relatively large (too large for most applications). Thus, there needs to be a correction for multiple testing that controls for the number of tests that are performed. Several different approaches exist to tackle this challenge, but typically

a stricter limit for individual significance tests is the procedure implicitly chosen.

Family wise error rate test corrections can be performed, such as the Bonferroni procedure (Aickin and Gensler, 1996) or the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). The false discovery rate (FDR) based procedure is another approach of controlling for multiple testing. If there is a likely follow-up test that can be performed to evaluate the significance of candidates identified, such as in the wet-lab through additional experiments, the FDR based procedure is typically the method of choice. While the family wise error rate test correction limits the probability of more than one false discovery, the false discovery rate is more akine to a limit on the probability that there is more than one false discovery (Shaffer, 1995). Therefore, the FDR approach has a higher chance of type I errors, but greater statistical power (Shaffer, 1995). The per-family error rate test correction is another option, which distinguishes between the precise number of false discoveries (Frane, 2015).

Time series

Often high-throughput experiments are not just performed for two contrasts of interest, such as stimulus and control, but over a specific time-frame. This allows the investigation of underlying dynamics in the dataset at hand. The specific sampling is generally conducted at several distinct time points, normally on the order of 5-10 due to funding constraints. The obtained data is termed longitudinal and specific statistical procedures exist to take advantage of this type of data. For example, a Gaussian process can be used to perform hypothesis testing in order to differentiate between the dynamics of two conditions over all obtained time points (Yang et al., 2016). Thus the order of time points is considered by the model. Another approach is the fitting of longitudinal data with splines and application of an F-test for differential dynamics. This procedure is advocated by the LIMMA package (Smyth, 2005). Lastly, the visualization, clustering and alignment of longitudinal data can also be performed by specific algorithms, such as fuzzy clustering (Futschik and Carlisle, 2005; Kumar and Futschik, 2007) or dynamic time warping (Aach and Church, 2001).

Gaussian Processes

A Gaussian process is a collection random variables of which any finite subset displays a multivariate Gaussian distribution (Rasmussen and Williams, 2006). In the framework of non-linear Bayesian regression the Gaussian process can be thought of as a prior distribution over functions completely specified by its mean $\mu(x)$ and co-variance or kernel function $K(x_i, x_j)$ (Kalaitzis and Lawrence, 2011).

$$f(x) \sim GP(\mu, K) \tag{4.1}$$

For applications that require specific smoothness assumptions a host of different kernel functions have been developed. Distributions not close to the standard Gaussian kernel can be used, such as a Matérn or periodic variants. The most common of these, which has as previously been applied to gene expression profiles, is the squared exponential kernel (Kirk and Stumpf, 2009). It is typically used in combination with a constant mean function. The characteristic length scale or smoothness parameter λ is left free and the amplitude parameter σ is set to unity in this case.

$$K_{SE}(x_i, x_j) = \sigma^2 exp(\frac{-(x_i - x_j)^2}{2\lambda^2}) \tag{4.2}$$

$$\mu(x) = 0 \tag{4.3}$$

Regression is then conducted by a Gaussian process on the time series by converting the prior over functions to a posterior. In addition, the smoothness parameter can be estimated from the data through a maximum likelihood approach, where commonly a nugget term is included to control for measurement errors (Pepelyshev, 2010). The trained process is subsequently used to predict values at locations previously not sampled.

$$f(y) \sim GP(\langle\mu\rangle, \langle K\rangle) \tag{4.4}$$

Overall, a Gaussian process can be used as an interpolation technique or to predict values of time points not yet sampled, such as before or after the measured values. Notably, in the field of geospatial analysis a similar technique is known as kirging.

Classification

The concept of classification is an important component of high-throughput data analysis. Often labeled or annotated samples exist due to prior information that can thus be used to train a classifier to predict the labels of new samples not yet annotated. The annotation step is typically performed by human experts and is overall a slow process. Therefore, classification algorithms are critical once the amount of data to be annotated increases past the manual annotation capacities. For the early diagnosis of disease, when it is not yet clear which are the important features for a diagnostic decision, classification algorithms can potentially outperform human experts by leveraging significantly more data, such as high-throughput databases to identify superior features.

On one hand, classification algorithms allow for the prediction of class labels of unknown samples as described, but on the other hand, a well functioning classifier can also be used in the validation of the importance of features, since a subset or convolution of features is typically selected by the algorithm to perform the classification. Hence, classification can be used as a validation technique in addition to

its common use in prediction. For example, if classification based on a particular subset of features performed better than classification based on another subset, then it stands to reason that the improvement stems from the particular features used. These features may be genes or proteins of interest. Such analyses can provide a first clue of potential causal relationships that are generally of importance in science. And, these clues can subsequently be investigated further by follow up experiments.

A wide range of classification algorithms have been developed in the past. The most simple variants often perform particularly well on large-scale real world datasets, as these do not suffer from overfitting if trained correctly. Standard classification algorithms include linear classifiers, such as Support Vector Machines (SVM) and tree based ensemble methods, such as Random Forrest (RF). The later have certain advantages and disadvantages depending on the dataset at hand. All such classification techniques fall in the realm of supervised learning techniques, because the data on which the algorithms are trained is annotated with class labels. Final predictions are always made on unlabeled data.

Support vector machines

In the case of SVMs a hyperplane or set of hyperplanes is constructed in the high dimensional space of the data, where a good separation between hyperplanes and training samples is the goal. Ideally, the functional margin is minimized (distance from the hyperplane to the closest training sample). Typically, this goes in hand with a lower generalization error of the trained classifier on the test dataset (Cristianini and Shawe-Taylor, 2000). If the dataset used for training is not linearly separable in the given dimensions, the problem is typically mapped into a higher dimensional space that allows for better separation. This is done with the use of kernel functions, which are selected to be appropriate for the problem (Press, 2007). Notably, the vectors defining the hyperplanes can be chosen to be specific linear combinations that facilitate computational efficiency.

Random forests

In the case of RFs a large number of decision trees is constructed for the test/train subsets of the data. Decision trees are simple linear classifiers that are invariant under scaling and can manage uninformative features (Friedman, Hastie, and Tibshirani, 2001). RFs correct for overfitting of individual decision trees and are also known as a form of ensemble learning, specifically bagging. Mechanistically RFs correct for the high variance produced by deep, but likely overfitted, decision trees through averaging of a large number of such trees. The interpretation of the resulting ensemble classifier is not as simple as in the case of decision trees, but still informative; for example, when understanding the importance of features is the goal of an investigation. Thus, RFs can be used as a method to obtain variable importance

information for the description of a particular dataset that can be used as first clue of potential causal relationships to be further investigated (Liaw and Wiener, 2002).

Cross-validation

For proper training and testing via cross-validation a dataset is typically split into a test and train set and in some cases an additional holdout set. The test set is where the parameters or hyper-parameters of an algorithm are optimized. A particular parameterization of the algorithm is then tested on the test set in order to evaluate its performance without overfitting, which would be the case if evaluation was performed on the train set. Thus, the fit of a particular model to a previously unseen dataset (test set) is evaluated. For certain applications, a holdout dataset is used for a final verification that can only be performed once. Importantly, the algorithm is not run again with different parameter settings. Thus, the holdout dataset provides protection against data dredging (or p-hacking) if used appropriately. The evaluation through cross-validation can be performed with different metrics. Most can be visualized in a diagram of a confusion matrix (Figure A.5). The constructed metrics consist of the number of true positives (T+), false positives (F+), false negatives (F-) and true negatives (T-). When true positives and false negatives are combined these yield the conditional positive cases (C+) or, alternatively, when the false positives and true negatives are combined these yield the conditional negative (C-) cases. False positives can be though of as a "false alarm" or type I error, while false negatives are comparable to a "miss" or type II error, true positives with a "hit" and true negatives with a "correct rejection" (Powers, 2011). Combining these elements yields more complex metrics commonly used during, cross-validation as shown below (Fawcett, 2006; Powers, 2011; Ting, 2011).

True positive rate

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{4.5}$$

True negative rate

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} \tag{4.6}$$

Positive predictive value

$$PPV = \frac{TP}{TP + FP} \tag{4.7}$$

Negative predictive value

$$NPV = \frac{TN}{TN + FN} \tag{4.8}$$

False negative rate

$$FNR = \frac{FN}{P} = \frac{FN}{FN + TP} = 1 - TPR \tag{4.9}$$

False positive rate

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - TNR \tag{4.10}$$

False discovery rate

$$FDR = \frac{FP}{FP + TP} = 1 - PPV \tag{4.11}$$

False omission rate

$$FOR = \frac{FN}{FN + TN} = 1 - NPV \tag{4.12}$$

Accuracy

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.13}$$

F1 score

An additional metric that uses a combination of the the outlined metrics in combination with binary counts is the F1 score. It is a compound metric equivalent in theory to the harmonic mean of precision and sensitivity (Powers, 2011) defined as,

$$F_1 = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN} \tag{4.14}$$

It is commonly used for the evaluation of binary classification algorithms, much as the receiver operating characteristic defined below. The F1 score can be interpreted as a weighted average of the PPV and TPR metrics, but instead scaled between zero and one. Notably, the F1 score does not take true negatives into account.

Receiver operating characteristic

Another important metric is the receiver operator characteristic (ROC). It is applied mostly to classification tasks where the performance of a binary predictor needs to be evaluated (Powers, 2011). For a ROC curve a single hyper-parameter (or threshold) of the classification algorithm is varied over a specified range to yield a sequence of ROC values forming the ROC curve. On the axes of the curve are true positive rate (TPR) contrasted with the false positive rate (FPR) (Powers, 2011). What is typically used as summary performance measure is the area under the ROC curve (AUC). However, the ROC metric does not consider the cost of false positives (F+) or false negatives (F−) and also not the underlying class distribution, which could be highly skewed. Therefore, it is only appropriate in homogeneously distributed, equally weighted settings.

Metrics

There exist various modifications of the metrics used in cross-validation (see Supplementary Figure A.5). For example, the F1 score can be modified for micro, macro or weighted averages. In the case of binary classification the false positive rate (FPR) and true negative rate (TNR) are often combined into misclassification rate, positive predictive value (PPV) or simply RMSE, if the values are continuously distributed (Powers, 2011). Another standard metric is the mean error and its derivatives, including the mean absolute error, mean squared error, mean squared log error and median absolute error. Further modifications include the described AUC as a component and the mean Silhouette coefficient, for the setting of clustering. Precisely how to choose an appropriate cross-validation metric depends on the goal of the particular analysis. In the classification of high-throughput data this goal is typically determined before an analysis is conducted in order to avoid overfitting. Overall, the metric most widely used is accuracy (ACC), as it is most intuitive to understand and simple to compute.

Sampling strategies

Cross-validation can be performed with different sampling strategies. Two of the most common strategies involve exhaustive and non-exhaustive cross-validation. Exhaustive strategies include leave-p-out cross-validation (LPOCV), which subsamples the dataset while randomly leaving out p of the samples in each sub-sampling, which are then used for test/train data split. Another method developed is k-fold cross-validation that randomly partitions the dataset into k separate subsets of which only one is used for testing and the remainder for training. In both cases this process is repeated until all the data has once been in the training and testing set (exhaustive cross-validation). Thus, full usage of the available data is possible without overfitting, as during each individual run a subset of the data is used and then the results of all subsets are averaged.

Stratification

Another important component of cross-validation is the stratification of samples. For certain machine learning techniques to be cross-validated, in particular classification, it is important that the number of samples per class are distributed equally. Otherwise, the train and test dataset are said to be imbalanced and the optimization of the chosen classifier will be highly biased towards performance for the samples of the predominant class. Another important issue is the potential existence of groups of samples, for example different samples from one particular group of cell types or patients that need to be considered when stratifying. In order to randomize the train and test dataset split appropriately in this scenario, the dependences likely existing

within the patient or cell type groupings need to be controlled. Therefore, it must be ensured that all of the groups found in the test dataset have at least one sample in the train dataset. Lastly, another type of dependence is found in time series or longitudinal data. Here, the groupings are ordered by time and it needs to be ensured that no samples in the train dataset fall temporally after those in the test dataset and vice versa.

Clustering

Clustering is an unsupervised statistical technique with the goal of detecting and visualizing patterns of a high-dimensional dataset. This technique is closely related to dimensional reduction, which is discussed in the Section 4.2.3. The most simple variant of clustering is hierarchical clustering. For example, agglomerative (bottom-up) schemes, such as Ward's method, or divisive (top-down) schemes, such as DIANA (Kaufman and Rousseeuw, 2009). Notably, the obtained dendrograms from hierarchical clustering are often highly dependent on hyper-parameter settings including the metric used. More complex approaches include spectral clustering (Han, Pei, and Kamber, 2011), affinity propagation (Frey and Dueck, 2007) and density based clustering approaches, such as DB-SCAN (Ester et al., 1996). The two most commonly used non-hierarchical clustering algorithms are the K-nearest neighbors technique (KNN) and K-Means clustering technique. The user defined constant K in both algorithms is a hyper-parameter that needs to be optimized appropriately via cross-validation and controls the number or nearest neighbors used or cluster centers initialized at.

Dimensional reduction

Dimensional reduction is often performed to visualize a dataset with a large number of features. Typically high-dimensional data can only be effectively depicted in 2D (maximally 3D) when visualized. Therefore, with the number of features obtained by high-throughput technologies that are on the order of 10-100 thousand, dimensional reduction is crucial to get a visual overview of the dataset at hand. Often, it is not possible to keep all relevant information during the dimensional reduction process. The data is typically obtained from complex systems, such as the cell, which inherently are not easy to reduce in dimensionality. Hence, either feature selection or feature extraction is performed during the dimensional reduction. Feature selection selects the 2-3 most interesting or important features for visualization for visualization (higher dimensions can not be displayed). However, all the information contained in the remaining non-displayed features is lost. Feature extraction aims to transform the high-dimensional measurements in such a way that the number of features is reduced without causing much loss of information. The most simple and widely used linear method for this purpose is principal component analysis (see Section 4.2.3). However, more cutting edges approaches, such as t-distributed stochastic

neighbor embedding (see Section 4.2.3), are also frequently employed. Many other techniques are based on non-linear methods and often do not scale well with the number of samples.

Principal component analysis

The dimensional reduction technique of principal component analysis (PCA) is preformed by finding principal components, e.g. eigenvectors, of the data. These components are ordered by the amount of variance they explain (see Supplementary Figure A.7). Thus, the top components can be used as a dimensional reduction that keeps most of the variation found in the data. However, these components are completely independent, e.g. orthogonal, to each other and thus uncorrelated. More formally, principal components are an uncorrelated orthogonal basis set (Wold, Esbensen, and Geladi, 1987). This allows for a computationally efficient algorithm, but unfortunately also constraints the types of components that can be found. PCA is sensitive to the relative scaling of features and samples depending on which dimension is reduced, but not to the absolute scaling of the data. Different algorithms exist to derive principal components, from those based on eigenvalue decomposition, to singular value decomposition and randomized techniques. Initially, the data is preprocessed to be mean centered and normalized to a standard deviations of unity. When PCA is applied it is used for exploratory analyses and data visualization to give a first overview of the data at hand. An important modification is kernel PCA (k-PCA), which uses a non-linear transform to map the data into a space that is more easily accessible to linear dimensional reduction techniques (Mika et al., 1999). This way data that is highly non-linear or complex, but simple to model linearly in the transformed domain, can be dimensionally reduced and visualized with the computationally efficient PCA technique.

t-SNE

Another approach at dimensional reduction that has gained popularity is termed t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton, 2008). This approach is limited by the computational complexity of its algorithm, especially when a large number of samples are used. However, the number of features can be on an order of magnitude consistent with most high-throughput data (without affecting the computational complexity). The t-SNE algorithm is found to produce very realistic reductions to low-dimensional space, even for the case of non-linear patterns. Therefore, it has been adopted as a method of choice for the visualization of small to medium sized high-throughput datasets. It uses the Kullback-Leibler divergence to optimize the similarity between the embedded (dimensionally reduced) data and the original data. The distributions that are thus optimized to be matching are based on pairs of points that when similar have a high-probability of being chosen and when dissimilar have a low-probability of being chosen. Within this scheme

the limited scalability to a large number of samples becomes apparent. The Student-t distribution is used to model the similarity distribution between points in the embedded data. Gradient descent is then applied as an optimization routine to find the locally optimal placement of points that minimize the Kullback-Leibler distribution between the embedded and the original similarity distribution.

Linear Models

Linear models are an important component of the statistical analysis of high-throughput data. Typically, in the framework of regression these models allow one to make predictions on future values of a response variable y from a number of observed explanatory variables x. The response variable is also termed a dependent variable and the explanatory variable an independent variable. A simple linear model is typical denoted as $y = \beta x$, where β is the parameter vector for fixed effects that is subsequently fitted.

There exist many extensions to this fundamental model. On one hand, generalized linear models (GLMs) have been developed to account for error distributions that are non-Gaussian or based on categorical data, while mixed effects models account for random instead of fixed effects. So termed generalized linear mixed models (GLMMs) can be effectively used when missing values are contained in the data. This is often the case in measurements obtained from high-throughput technologies. On the other hand, modifications also exist of linear models which correct effects due to heteroscedasticity or errors-in-variables. These modification explicitly correct for measurement errors in both the dependent and independent variables and are also termed Deming regression or orthogonal regression. If the variable y or x in the above linear equation is a vector (or matrix), then multiple (or multivariable) regression and general (or multivariate) regression are appropriate. In the case of too many response variables y and too few samples x measured, Tikhonov regularization or ridge regression is often appropriate, which can deal with underdetermined systems of equations and extends the standard ordinary-least-squares (OLS) based fitting approaches to this common setting.

Model selection

According to Occam's razor[3] a theory or model should not be more complex than necessary. Such a view leads to a preference for simpler models, which are more testable due to falsification being more likely if the model is more complex. For the case of model selection, this principle is made explicit by the use of the Bayesian information criterion (BIC) or the Akaike information criterion (AIC). Importantly, if one needs to distinguish between two equally accurate models, these measures can be used as an aid or diagnostic.

[3] An important heuristics in science.

Bayesian information criterion

The BIC allows for the selection of models among a finite subset thereof. The lower the BIC score, the more preferred is the model that is tested. The corresponding formula is given as,

$$BIC = \ln(n)k - 2\ln(L_{max}) \tag{4.15}$$

where k is the number of parameters of the respective model analyzed, L_{max} is the maximized value of the likelihood function of this model and n is the number of samples x obtained to validate the model (Wit, Heuvel, and Romeijn, 2012). Equation 4.15 is roughly based on the likelihood function, which is improved by adding additional parameters to a model. However, BIC penalizes these additional parameters in the above fashion. This guards a model against overfitting the dataset x at hand, which can lead to inferior predictions on potential new data.

Akaike information criterion

The AIC is in contrast not based on the number of samples or data points under consideration. It penalizes the overall number of parameters less strongly than BIC and exhibits a theoretical advantage, as it can be derived from principles of information theory. The corresponding formula is given as,

$$AIC = 2k - 2\ln(\hat{L}) \tag{4.16}$$

where $\hat{L}$ is the maximized value of the likelihood function of the model as above and k is the number of parameters of the model (Akaike, 1974).

4.3 Results

The workflows outlined in the preceding section were applied to analyze and interpret the high-throughput experiments conducted in the framework of the LungSys and GerontoSys consortia. Especially workflows aimed at translating quantitative measurements into categorical information are highlighted in the following sections. These workflows have provided important feedback to research collaborators in the wet-lab and clinic and promote the integrative research focus at the core of the two consortia (see Section 4.1). The contributions to the consortia with respect to the developed workflows have lead to several publications in peer-reviewed journals (Ohse, 2018; Kalfalah et al., 2015; Kalfalah et al., 2014). In the following sections a case study of each consortium is discussed in detail.

4.3.1 Consortium: LungSys

The focus of the LungSys consortium is the investigation of therapy resistance in non-small cell lung cancer, which makes up about 80% of lung cancer cases world wide (LungSys Consortium, 2017). The typical cause of death in these cases is from a metastatic spread of cancer cells and subsequent major organ failure of the patient (LungSys Consortium, 2017). Often, the metastatic process has already started by the time lung cancer is diagnosed and therapeutic options are then generally limited to chemotherapy or in some cases tyrosine kinase inhibitors. When therapy resistance develops, the tumor continues to grow despite medical interventions. This complex disease progression in non-small cell lung cancer spans different spatial and temporal scales of molecular and cell biology. Thus, it is crucial to obtain measurements in a global and dynamical fashion in order to understand the development of therapy resistance.

The systems biological approach of the LungSys consortium places particular emphasis on the characterization of the mechanisms that facilitate the early metastatic development of lung cancer in order to yield insights that might be used for improved early diagnosis and treatment (LungSys Consortium, 2017). Mechanisms underlying early metastatic development involve the separation of cancer cells from the primary tumor site, subsequent invasion of the surrounding tissues, entry into the blood stream and terminally the invasion of other organs and tissues (LungSys Consortium, 2017). At various stages of this process therapy resistance can develop. Of specific interest here is an in-depth investigation of already characterized factors involved in therapy resistance. Particularly the effects of hepatocyte growth factor, transforming growth factor β, insulin-like growth factor and the erythropoietin receptor are studied in the LungSys consortium across different spatial and temporal scales by the use of high-throughput technologies (LungSys Consortium, 2017). Most of the material highlighted in the following case study is taken verbatim or in modified form from publication C.1 (Ohse, 2018).

4.3.2 Case study 1

Through the application of the developed workflows and in collaboration with wet-lab experiments performed at the DKFZ partner site and clinical studies at the university hospital Heidelberg, it was possible to determine a prognostic and predictive biomarker for squamous cell carcinoma of the lung (LUSC) that could serve as a potential molecular target for early clinical intervention. Specifically, the expression ratio of MYO10 was identified to be prognostic and predictive for overall survival of squamous cell lung cancer. Its role in motility and invasion was quantified and validated by 2D migration and 3D invasion assays of the TGFβ-stimulated squamous cell lung cancer cell line SK-MES1 (see Figure 4.2) and relevant candidate genes in the context of squamous cell lung cancer were identified through time resolved high-throughput experiments with next generation sequencing (RNA-seq) and microarray technology (see Figure 4.3 to Figure 4.6). The particular expression of MYO10 was then assessed in a 151-patient squamous cell lung cancer patient cohort of paired surgically-resected tissues with 80-month clinical follow-up.

Adenocarcinoma of the lung (LUAD) and squamous cell carcinoma of the lung are the two major subtypes of NSCLC. Although the prevalence of LUSC in developed countries is declining, it still accounts for about 25% of NSCLC cases (Drilon et al., 2012). Despite the progress in developing targeted approaches in LUAD, therapeutic options for LUSC remain very limited as driver oncogene mutations are uncommon (Rooney, Devarakonda, and Govindan, 2013). Platinum-based chemotherapy has been the gold standard for first-line therapy for LUSC patients. However, in a significant proportion of patients cancer cells are resistant to chemotherapy and the disease rapidly progresses (Kim et al., 2008). Thus, there is an urgent need to gain insights into the mechanisms contributing to LUSC in order to establish biomarkers that help clinicians identify patients at the highest risk for disease progression and therapy resistance.

Both early metastasis and therapy resistance are attributed to cancer cells undergoing epithelial-to-mesenchymal transition (EMT) and acquiring a more invasive phenotype with cancer stem cell-like properties (Yu et al., 2013). Tumor cells harboring EMT features were repeatedly reported to localize on the invasive front of the tumor, hence mediating cancer cell dissemination and metastasis (Maeng et al., 2014). There is growing evidence that deregulated TGFβ signaling contributes to the acquisition of an EMT phenotype by lung cancer cells. In the context of LUSC, elevated TGFβ1 levels were correlated with poor patient prognosis (Sterlacci et al., 2012) and over-activation of the TGFβ pathway was reported as a common feature in lung cancer (Marwitz et al., 2016). Moreover, the EMT phenotype was widely observed in surgically resected specimens and associated with a worse clinical outcome and chemoresistance (Shintani et al., 2011). However, a mechanistic understanding of

TGFβ-induced changes and their impact on LUSC progression remains to be established. Therefore, phenotypic and transcriptome wide approaches were combined in this study to determine TGFβ-induced dynamic changes in the transcriptome of a LUSC cell line. Thereby a candidate prognostic biomarker was derived and subsequently validated in a clinical cohort.

TGFβ treatment enhances pro-tumorigenic properties of LUSC cells

To study the impact of TGFβ on LUSC cells the cell line SK-MES1 was used as a model system. By quantitative immunoblotting it was shown that TGFβ-induced phosphorylation of Smad2 and Smad3 in SK-MES1 cells reached a maximum at 30 minutes and afterwards declined. SK-MES1 cells usually grow in tight epithelial colonies, but after treatment with TGFβ cell-cell contacts were lost and an elongated spindle-shaped morphology was acquired (Figure 4.1). This is a feature commonly observed upon TGFβ-induced epithelial-to-mesenchymal transition (EMT). In line with these morphological alterations, TGFβ treatment of SK-MES1 cells induced the mRNA expression of classical EMT markers such as SNAI1, ZEB1, VIM and MMP9.

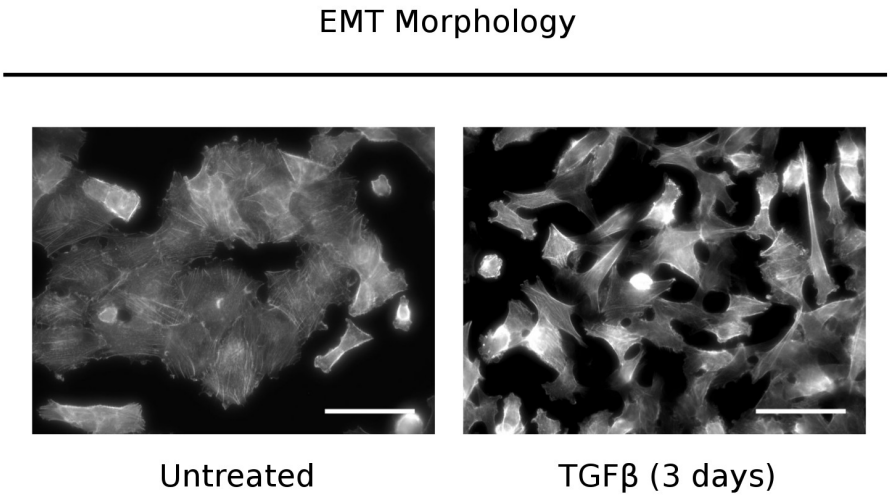


FIGURE 4.1: Prolonged exposure of SK-MES1 cells to TGFβ1 induces acquisition of EMT-like morphology. Cells were either stimulated with 2 ng/ml TGFβ1 or left untreated for 3 days, fixed and stained for F-actin (white) and DNA (solid gray). Scale bar corresponds to 50 μm (Ohse, 2018).

The activation of TGFβ signal transduction and target gene expression as well as the morphological changes were explored. Therefore, workflows were established to quantitatively assess at the single cell level the impact of TGFβ on SK-MES1 cells in a 2D cell migration assay and a 3D collagen invasion assay (Figure 4.2). In the 2D migration assay it was observed by analyzing more than 1000 of single cell tracks per

condition that the TGFβ treatment resulted in a two-fold increase in migration speed (from 4 to 8μm/h). Co-treatment with a type I TGFβ receptor inhibitor prevented this effect. In the 3D collagen invasion assay TGFβ treatment resulted in a two-fold increase in collagen-invaded SK-MES1 cells. Some of the TGFβ-treated SK-MES1 cells invaded more than 100 μm into the dense collagen gels, while untreated cells invaded on average not more than 20 μm (Figure 4.2). The increase in the invasion capacity was TGFβ-specific because it was abolished by co-treatment with a type I TGFβ receptor inhibitor.

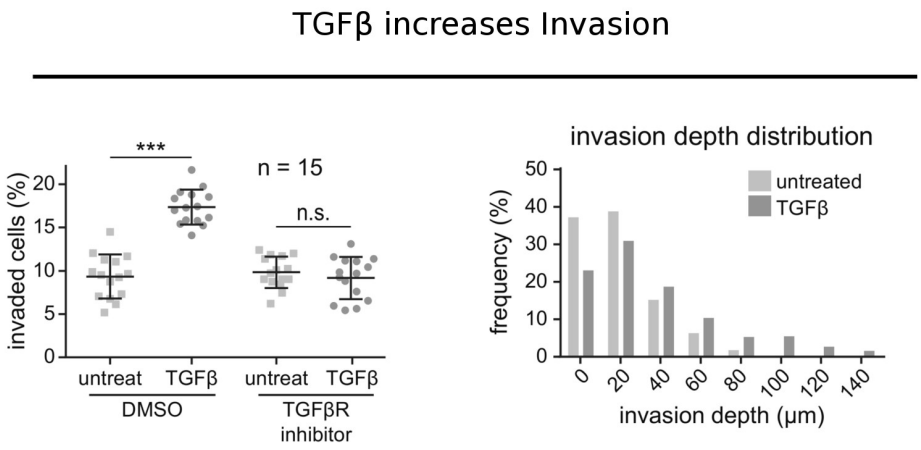


FIGURE 4.2: TGFβ stimulation increases number of invading cells and the average invasion depth. SK-MES1 cells were seeded in 96-well plates with precast collagen gels, allowed to attach overnight, growth factor-depleted for three hours, pretreated with either SB-431542 or DMSO, stimulated with 2 ng/ml TGFβ1, allowed to invade for four days, stained with Hoechst and imaged with a confocal microscope. The number of invaded cells and invasion depth were assessed. One representative experiment is shown. Data are presented as median and SD, every dot corresponds to a biological replicate (n = 15). N indicates the number of invaded cells (Ohse, 2018).

It was reported that the EMT phenotype correlates with increased resistance to chemotherapy (Arumugam et al., 2009). To examine the impact of TGFβ on the resistance of SK-MES1 cells to cisplatin, a cell viability assay based on metabolic activity and an apoptosis assay based on caspase 3/7 activity was employed. It was observed that pre-treatment of SK-MES1 with TGFβ for 3 days resulted in a 4.4-fold increase of viable cells after 3 days exposure to 10 μg/ml cisplatin. Likewise, pretreatment with TGFβ reduced the caspase 3/7 activity across all tested doses of cisplatin by 25%. Collectively this indicate that SK-MES1 cells acquire a more aggressive phenotype upon exposure to TGFβ.

Multiple actin cytoskeleton and motility related genes are up-regulated in LUSC cells upon TGFβ stimulation

To elucidate mechanisms that contribute to TGFβ-induced tumor spread and chemotherapy resistance in LUSC, a time-resolved whole-transcriptome RNA-Seq analysis of SK-MES1 cells was performed. The cells were treated with TGFβ for up to 48 hours or were left untreated, e.g. as control. Genes were considered as differentially regulated if their overall mRNA expression dynamics in treated versus untreated cells was significantly different (multiple testing adjusted p-value <0.01). The resulting list of differentially regulated genes was used for gene set enrichment (see Section 4.2.3) to identify regulated gene ontology categories of cellular components, which were subsequently visualized with the REVIGO tool (Supek et al., 2011) to establish clusters with distinct gene expression patterns. See also Supplementary Figure A.6. This approach revealed a preferential regulation of four gene clusters encoding actin cytoskeleton-, motility-, ECM- and secretory-related proteins (Figure 4.3). Through the application of the measure of biological relevance algorithm (see Chapter 3) it was possible to exclude differentially regulated but highly general GO categories from the initial exploratory analysis that was performed. This led to a better understanding of the cellular changes specific to TGFβ stimulation. To narrow down the list of potential candidates involved in mediating the TGFβ-induced invasive properties of LUSC cells, the five genes per cluster with the lowest multiple testing-adjusted p-values and with at least two-fold peak amplitude change after normalization to corresponding untreated samples were selected. This resulted in a list of 15 TGFβ regulated genes because some of the genes were among the top five candidates in more than one cluster (Figure 4.4). Interestingly, genes identified as candidates in the approach included MYO10, SERPINE1, ITGB3, ITGA5, TGFBI, VIM and MARCKS. These TGFβ regulated genes were previously associated with increased cancer invasiveness, chemoresistance and worse clinical outcome in different cancer entities including breast, NSCLC, invasive melanoma and prostate cancers (Makowska et al., 2015; Garibay et al., 2015; Lauden et al., 2014; Look et al., 2002; Haider et al., 2014; Chen et al., 2017). To determine which of these TGFβ regulated genes are relevant in the context of LUSC, the alterations of mRNA levels of the selected 15 candidate genes were evaluated in a cohort of 501 LUSC patients (Figure 4.5) from the Cancer Genome Atlas (TCGA). Strikingly, the MYO10 gene was up-regulated in 27% of the patients, whereas an up-regulation of the mRNAs of the other genes was only observed in 2-5% of the LUSC patients. Interestingly, the genes with the highest percentage of mRNA up-regulation in LUSC patients belonged either to the migration or the actin cytoskeleton clusters, while genes from the ECM and secretory clusters were rarely altered in LUSC patients, although several of the genes from these clusters showed a high fold increase in SK-MES1 cells upon TGFβ treatment (Figure 4.4). Given the prominent up-regulation of MYO10 expression in LUSC patients and the pivotal role of non-muscle myosins in mediating cancer cell

invasion in multiple cancer entities (Ouderkirk and Krendel, 2014), it was examined whether other myosin-encoding genes scored high in the analysis but were not among the top five regulated genes. Indeed, second and third most regulated genes encoding myosins were MYH9 and MYO1E, which were previously implicated in cancer progression (Hallett et al., 2012; Katono et al., 2015). Both of these myosin genes were up-regulated in LUSC patients of the TCGA cohort with MYH9 being overexpressed in 7% of the cases (Figure 4.5 lower panel). Furthermore, a significant co-occurrence of an up-regulation of the mRNAs of MYO10, MYH9, MYO1E and TGFB1 was observed in LUSC patients of the TCGA cohort, suggesting that the exposure of tumor cells to elevated levels of TGFβ might have stimulated up-regulation of motility and invasion-related myosins. Therefore, all three myosin genes were included for further analysis.

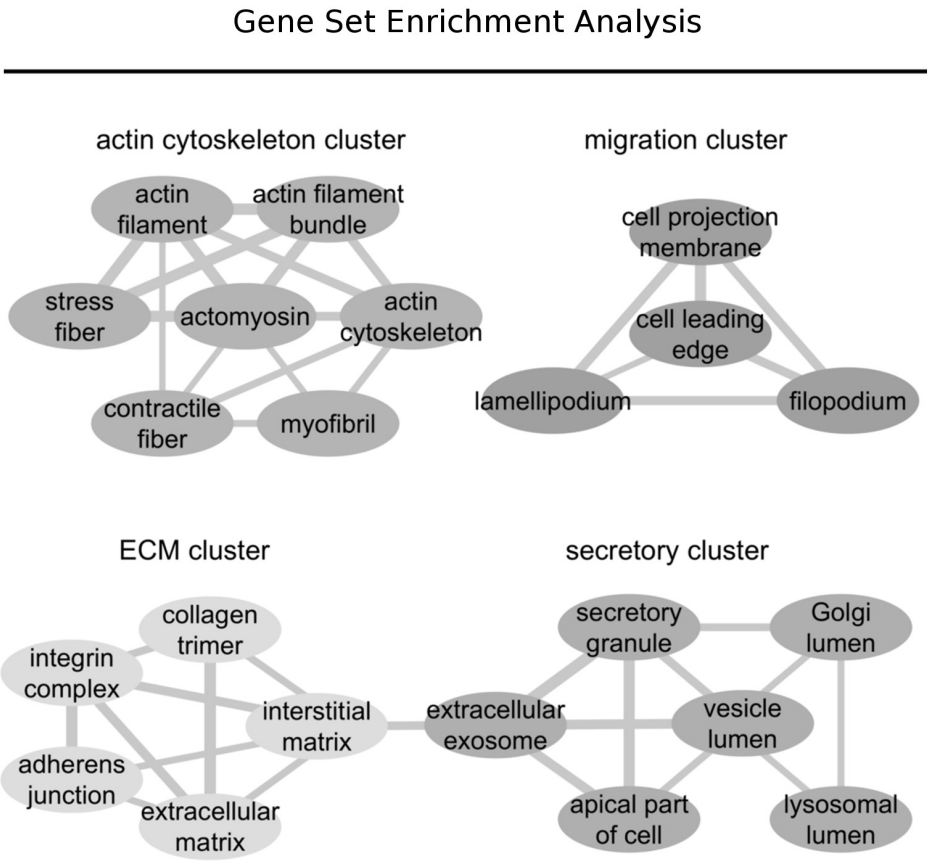


FIGURE 4.3: TGFβ treatment results in up-regulation of actin cytoskeleton and motility related genes. Shown are clusters of significantly up-regulated GO cellular component based gene sets between TGFβ-treated and untreated conditions. SK-MES1 cells were stimulated with 2 ng/ml TGFβ or left untreated. Significantly up-regulated GO categories were visualized using REVIGO (threshold for up-regulation $p < 0.01$, allowed similarity 0.5). Thickness of connecting gray lines corresponds to the similarity of the GO categories. Only clusters that consist of at least two GO categories are displayed (Ohse, 2018).

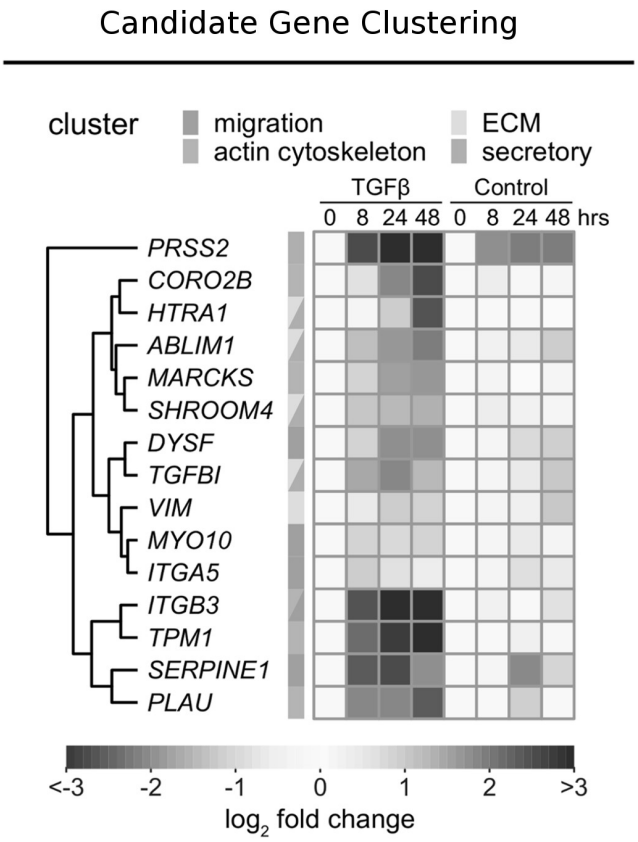


FIGURE 4.4: Time-resolved dynamics of top differentially regulated candidate genes from each of the clusters. Top five genes from each of the four clusters with the lowest adjusted p-values and fold change of at least two after normalization to untreated samples were selected as candidates. In case the same gene belonged to different clusters and satisfied the inclusion criteria, it was marked as belonging to both clusters (Ohse, 2018).

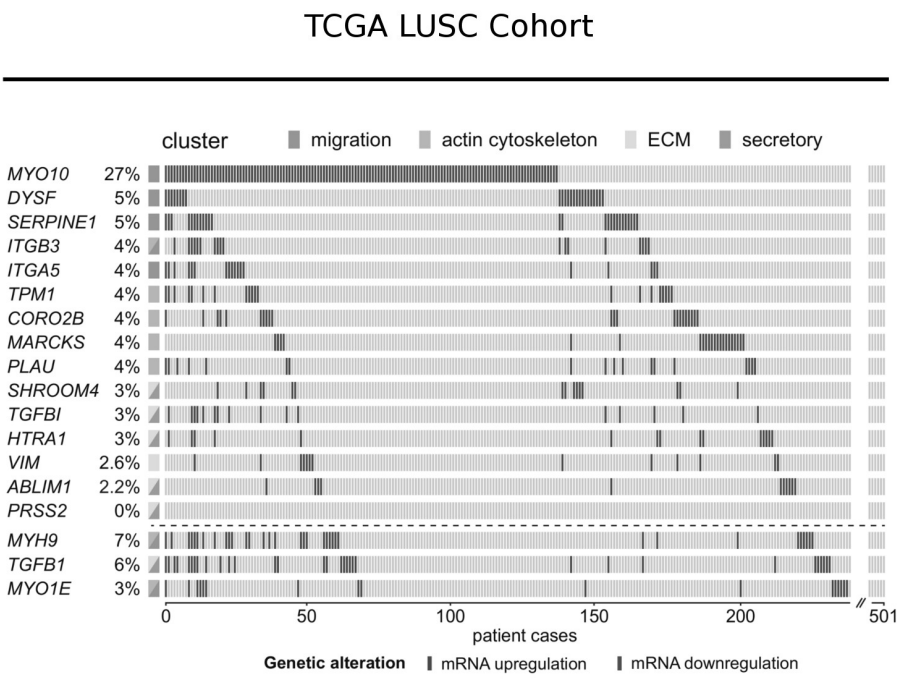


FIGURE 4.5: TCGA LUSC cohort of RNA-Seq expression data of selected candidate genes sorted by frequency of mRNA up-regulation. The genes MYH9, TGFB1 and MYO1E were additionally included (Ohse, 2018).

TGFβ-induced myosin motors are essential for TGFβ-mediated cancer cell invasion

To examine the biological importance of the candidate TGFβ-induced myosins RNA-Seq data (Figure 4.6) was validated with time-resolved samplings by qRT-PCR of TGFβ-stimulated SK-MES1 cells. See also Supplementary Figure A.8. In line with the dynamics of gene expression observed by RNA-Seq, all three candidate genes demonstrated strong mRNA induction upon TGFβ treatment, with MYO10 expression being the most pronounced and sustained. Given the role of non-muscle myosins in cancer metastasis, the effect of gene silencing on the ability of the SK-MES1 cells to invade 3D collagen gels was studied in response to TGFβ stimulation. To knockdown MYO10, MYH9 or MYO1E, siRNA was used and achieved a knockdown efficiency of more than 85% at the mRNA level. Whereas TGFβ treatment of SK-MES1 cells transfected with control non-targeting siRNA resulted in a two-fold increase in the number of invaded cells, the TGFβ-enhanced invasion of SK-MES1 cells was abrogated upon down-regulation of the different myosins. These results indicate that the TGFβ-induced non-muscle myosins MYO10, MYH9 and MYO1E play a non-redundant and crucial role in mediating invasiveness of the LUSC cells.

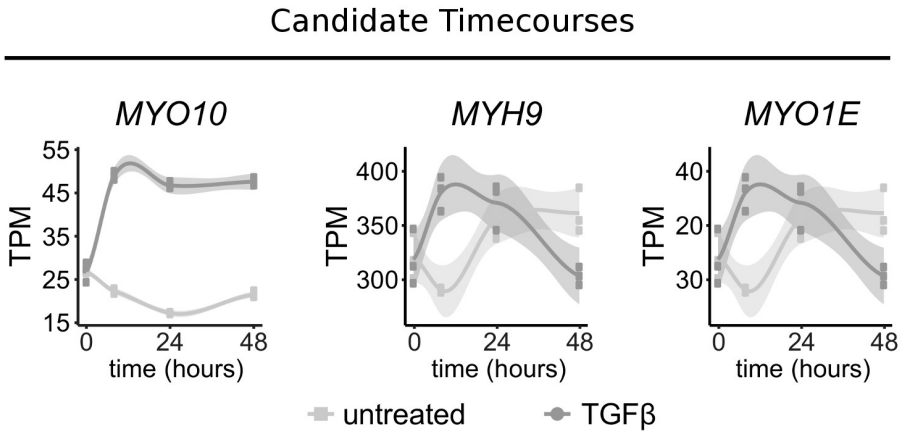


FIGURE 4.6: Time-resolved transcriptome data of selected myosin genes upon TGFβ treatment in SK-MES1 cells. Cells were growth factor-depleted for three hours and stimulated with 2 ng/ml TGFβ1 or left untreated. The mRNA was extracted and sequenced using HiSeq 4000. Data is presented in TPM (transcripts per million) values. Each dot represents a biological replicate, shaded areas correspond to standard errors (Ohse, 2018).

MYO10 mRNA overexpression is prognostic for overall survival of patients

Actin-based protrusions and TGFβ-induced myosins are crucial for multiple phases of the metastatic cascade (Ouderkirk and Krendel, 2014). Among identified TGFβ regulated non-muscle myosins MYO10 gene showed the strongest up-regulation in response to TGFβ stimulation in SK-MES1 cells and the highest mRNA overexpression in LUSC patients of the TCGA cohort. Therefore, its clinical relevance in paired tumor and tumor-free tissue from the NSCLC cohort consisting of both LUAD and LUSC patients was further assessed. For each tumor entity, patients were divided into two subgroups based on the expression ratio of MYO10 mRNA in tumor versus tumor-free tissue, MYO10 fold change <1 and MYO10 fold change >1, respectively. To investigate the prognostic value of the MYO10 mRNA expression ratio, Cox regression analysis was performed. Univariate analysis indicated that a high MYO10 mRNA expression ratio, gender and the pathological stages were prognostic factors for the overall patient survival. The multivariate analysis suggested that a high MYO10 mRNA expression ratio was only prognostic for LUSC patients, but not for LUAD patients. Using the MYO10 mRNA expression ratio to separate the patient groups, it was confirmed that LUSC patients with high MYO10 mRNA expression ratio demonstrate reduced overall survival (P = 0.008), which was not observed in LUAD patients (P = 0.57). Next, the LUSC patients were subdivided according to those that had no further treatment after the resection and those who received adjuvant chemotherapy (Figure 4.7). This analysis showed that for untreated patients

MYO10 mRNA expression ratio expression was not predictive for overall survival (P = 0.429). On the contrary, patients with low MYO10 mRNA expression ratio strongly benefited from the adjuvant chemotherapy treatment in comparison to patients with a high MYO10 mRNA expression ratio (P = 0.429). Therefore, it can be concluded that MYO10 mRNA expression ratio is predictive for the outcome of adjuvant chemotherapy treatment of LUSC patients.

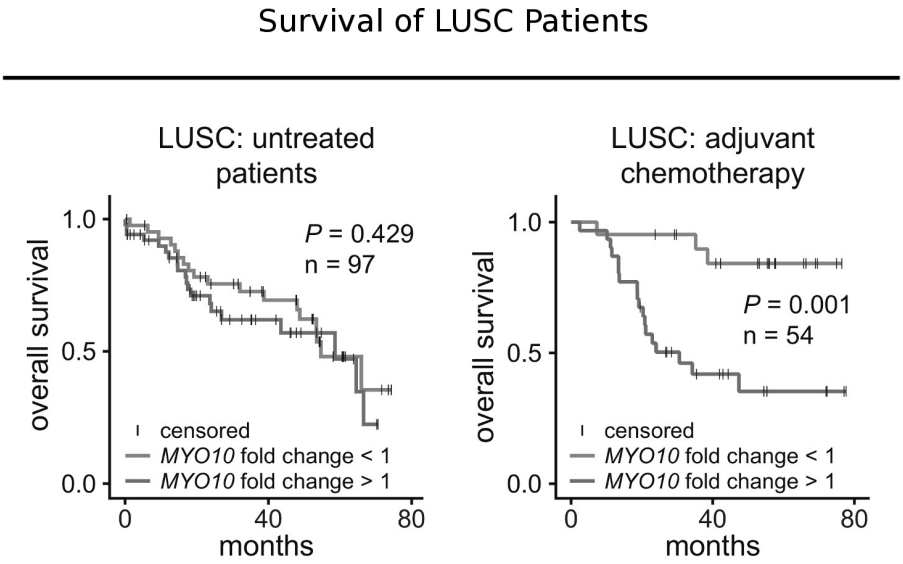


FIGURE 4.7: Kaplan-Meier curves for adjuvant chemotherapy response in MYO10 low (left) and MYO10 high (right) patients. Significance of difference between the two groups was tested using non-parametric Wilcoxon-Mann-Whitney test (Ohse, 2018).

Because of the observed enhanced chemoresistance of LUSC cells after TGFβ treatment and because TGFβ-induced EMT has been associated with chemotherapy resistance in patients (Shintani et al., 2011; Soltermann et al., 2008), the expression of EMT markers in tissue of LUSC patients was determined. Notably, patients with an elevated MYO10 mRNA expression ratio displayed a higher expression of EMT signature genes such as SNAI2, TWIST1 and VIM. The fact that TGFβ is one of the most potent EMT-inducers (Lamouille, Xu, and Derynck, 2014) and the co-occurrence of MYO10 and TGFB1 mRNA up-regulation in a substantial proportion of LUSC patients suggest that activation of TGFβ signaling might trigger the observed alterations in LUSC patients. Finally, a higher MYO10 mRNA expression ratio was observed in patients with stage III disease, making it prognostic for patients with a higher pathological stage and affected local or distant lymph nodes (Figure 4.8). Taken together, the here presented work suggest that the mRNA expression ratio of MYO10 can be used as a new independent prognostic biomarker for survival in

patients with resected LUSC.

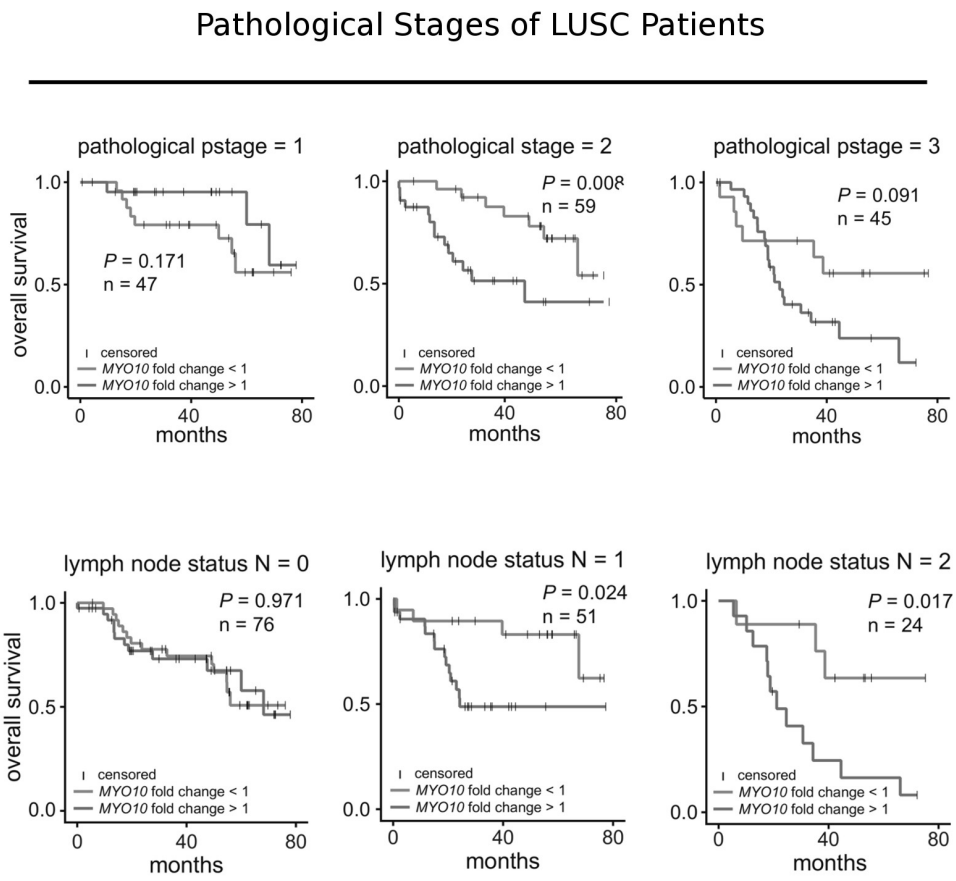


FIGURE 4.8: Kaplan-Meier curve for the pathological stages of LUSC patients and lymph node status using MYO10 mRNA expression ratio. Significance of difference between the two groups was tested using non-parametric Wilcoxon-Mann-Whitney test (Ohse, 2018).

In summary, it was determined that the TGFβ stimulation of SK-MES1 cells resulted in an increase of cancer cell invasion and cisplatin resistance. The global high-throughput analysis of the expression dynamics of TGFβ-induced genes revealed an up-regulation of networks of motility and actin cytoskeleton related genes and of these MYO10 was most prominent. Subsequently, it was demonstrate that squamous cell lung cancer patients with a high expression ratio of MYO10 in resected tumors versus tumor free tissue showed a lower overall survival and responded less to adjuvant chemotherapy. Hence, MYO10 represents a new prognostic and predictive biomarker for squamous cell lung cancer and due to its role in motility and invasion could serve as a molecular target for therapeutic interventions in patients with this aggressive disease. The high-throughput workflows described in Section 4.2 were an integral part of the conducted analysis, especially during the early exploratory

stage.

4.3.3 Consortium: GerontoSys

The focus of the GerontoSys consortium is the interdisciplinary study of human aging. Neurological, cardiovascular and other systemic diseases are all closely associated with age (GerontoSys Consortium, 2017). This specifically includes dementia, heat attacks and cancer that are some of the diseases most prevalent in aging populations. Hence, it is important to understand the underlying biological processes and their specific impact on disease development (GerontoSys Consortium, 2017). By investigating such underlying processes a foundation for the study of prevention and early detection of age associated disease is envisioned. The model system explored by the consortium is the human skin, due to ready availability of biological samples from this organ and the well established methods to culture and assess the phenotype of skin derived cells. Experiments are performed with the aid of high-throughput technology, clinical surveys and in conjunction with additional experiments on the molecular level that investigate some of the specific mechanisms identified (GerontoSys Consortium, 2017). The overall consortium strategy is to investigate different levels of cellular changes in the skin that are inherent to aging and to integrate the hereby obtained high-throughput data in a holistic fashion. Most of the material highlighted in the following case study is described in detail in publications B.1 and B.2 (Kalfalah et al., 2015; Kalfalah et al., 2014).

4.3.4 Case study 2

In the first set of experiments performed within the GerontoSys consortium, the focus has been on the analysis of fibroblast cells isolated from the skin of female human donors. Structural chromosome abnormalities were subsequently identified and increased DNA strand breaks or repair deficiencies were analyzed in detail. In order to address the specific hypothesis that skin cell aging can lead to chromosome instability, fibroblasts cells provided a good model system, because of their exposure to environmental factors including the sun in addition to the possibility to isolate fibroblasts relatively easily via a biopsy.

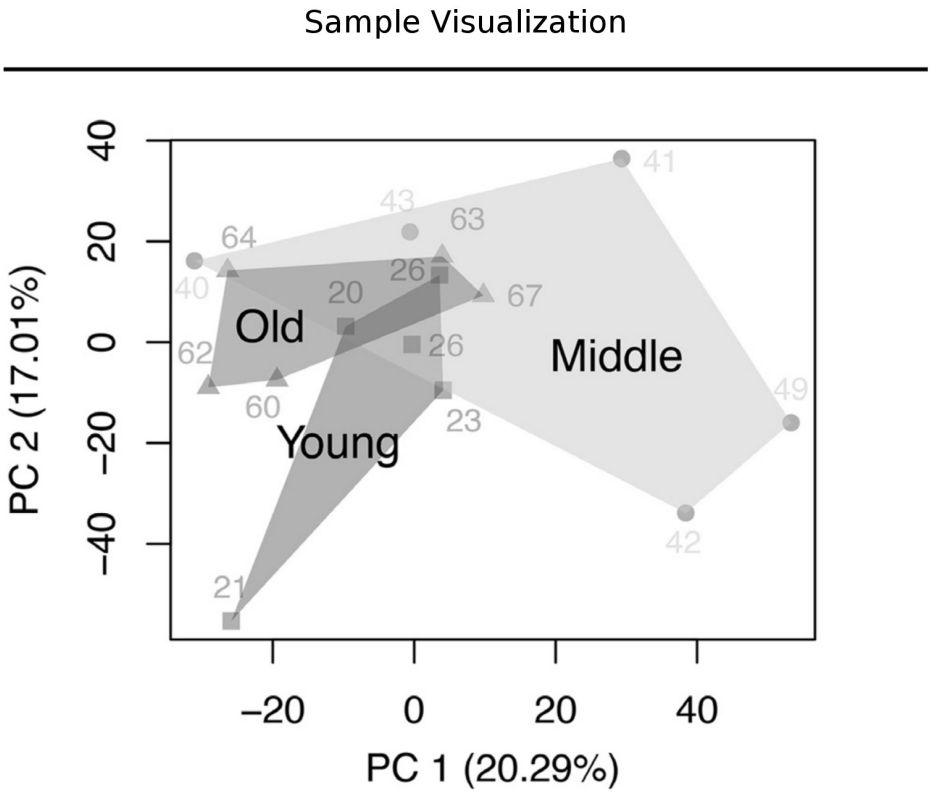


FIGURE 4.9: Principal component analysis based visualization of patient samples from different age-groups to obtain an overview of the different transcriptome profiles. Samples consisting of human dermal fibroblasts of the same patient age group are enclosed by a convex hull to mark the overlap and separation of age groups (Kalfalah et al., 2015).

Apart from the high-throughput analysis of gene expression measurements according to the workflows described in Section 4.2, the telomere length of human dermal fibroblasts was characterized, but found insignificant with respect to the investigated hypothesis. An overview of high-throughput measurements is given in Figure 4.9 by the PCA method. Furthermore, gene set enrichment analysis was performed to reveal impairment of mitosis, maintenance of telomeres and chromosomes and the induction of genes related to DNA repair and non-homologous end-joining, of which specifically XRCC4 and ligase 4 were of greatest interest (see Figure 4.10). With increasing age the proliferation rate of cells dropped and heterochromatin marks, structural chromosome abnormalities, DNA strand breaks and histone phosphorylation (H2AX) increased.

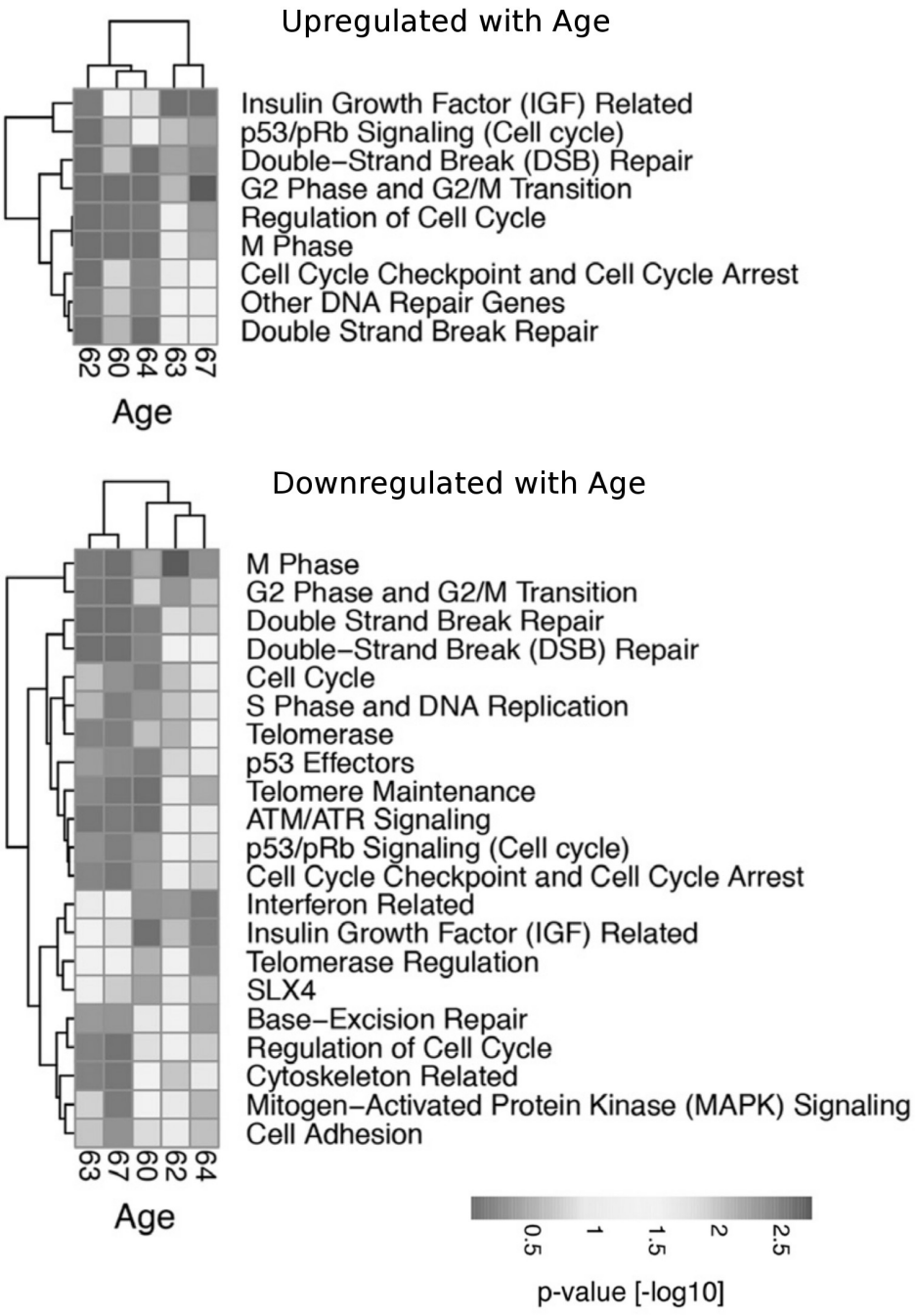


FIGURE 4.10: Gene set enrichment analysis analysis of age-related regulation of genes associated with genome maintenance. The two heatmaps depict the enrichment analysis of gene sets related to cell cycle, senescence, telomere and DNA repair, which are up- or down-regulated with age. Heatmap greyscale values correspond to the -log10 transformed p-values. Depicted expression values are row-wise mean centered and scaled to unit variance. Genes and samples have been hierarchically clustered using complete linkage (Kalfalah et al., 2015).

However, in a large fraction of cells from older donors the repair of DNA strand breaks induced via X-rays was reduced, even though DNA repair genes were up-regulated. The observed phenotype of genome instability, increased heterochromatinisation and continued up-regulation of DNA repair genes (without an effect in a large fraction of the fibroblast cells) indicated that the overall phenotype is not one of senescence but one of aging related changes. Specifically, because proliferation was observed to be stable.

In the next sequence of experiments the focus was again on the analysis of skin cells isolated from human donors. Here, the dermis was a focus of experiments. The cells in the dermis are post-mitotic and connected by an extracellular matrix to form an important layer of the skin. These cells in particular are prone to age associated damage accumulation and abnormal changes in the form of mitochondiral and nuclear dysfunctions. Such potential causes of dermal aging were further studied by isolating fibroblasts from human donors of different age groups and culturing these cells at low population doubling times to preclude the effects of replicative senescence. An overview of the acquired high-throughput data was again performed with the PCA method. The most prominent finding of transcriptome analyses with respect to aging was the decreased expression of mitochondrial genes. Consistent with these findings mitochondrial content and cell proliferation diminished in cells from tissue donors with increasing age. Important associated factors observed to be up-regulated were AMP-dependent protein kinase (AMPK), PPARγ-coactivator 1α (PGC1A) and down-regulation of NAD+-dependent deacetylase sirtuin 1. See also Figure 4.11 and Figure 4.12. Specifically, for the cells derived from older donors, the PGC1A-mediated mito-biogenetic response to direct AMPK-stimulation by AICAR was unaffected. At the same time, the PGC1A-independent mito-biogenetic response to starvation diminished in addition to increased ROS production. Thus, a decline in PGC1A-independent mito-biogenesis is likely, which is not appropriately compensated by changes in the AMPK/PGC1A pathway. This leads to decreased mitochondrial content and thus a reductive overload of residual respiratory capacity (Kalfalah et al., 2014). Overall, it was found that inadequate mito-biogenesis in primary dermal fibroblasts from old humans is associated with impairment of PGC1A-independent stimulation (Kalfalah et al., 2014).

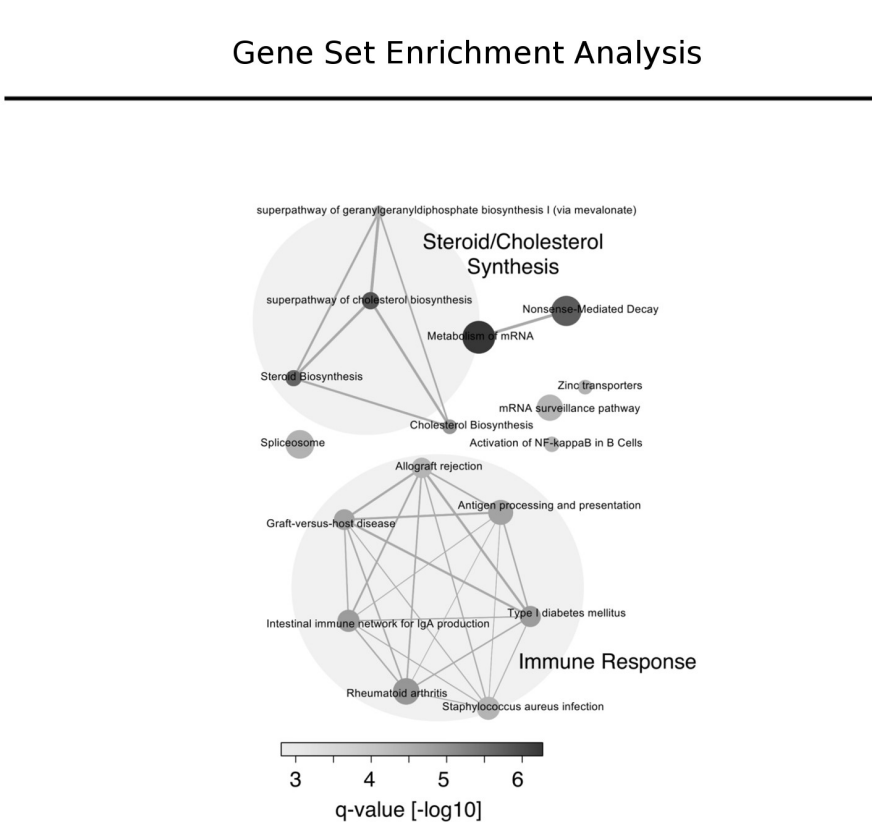


FIGURE 4.11: Gene set enrichment analysis of young and old fibroblast samples using gene sets from the Consensus Path DB (Kamburov et al., 2008). Shown are significantly up-regulated gene sets with age. Size and greyscale of the nodes are proportional to the gene set size and significance (Kalfalah et al., 2014).

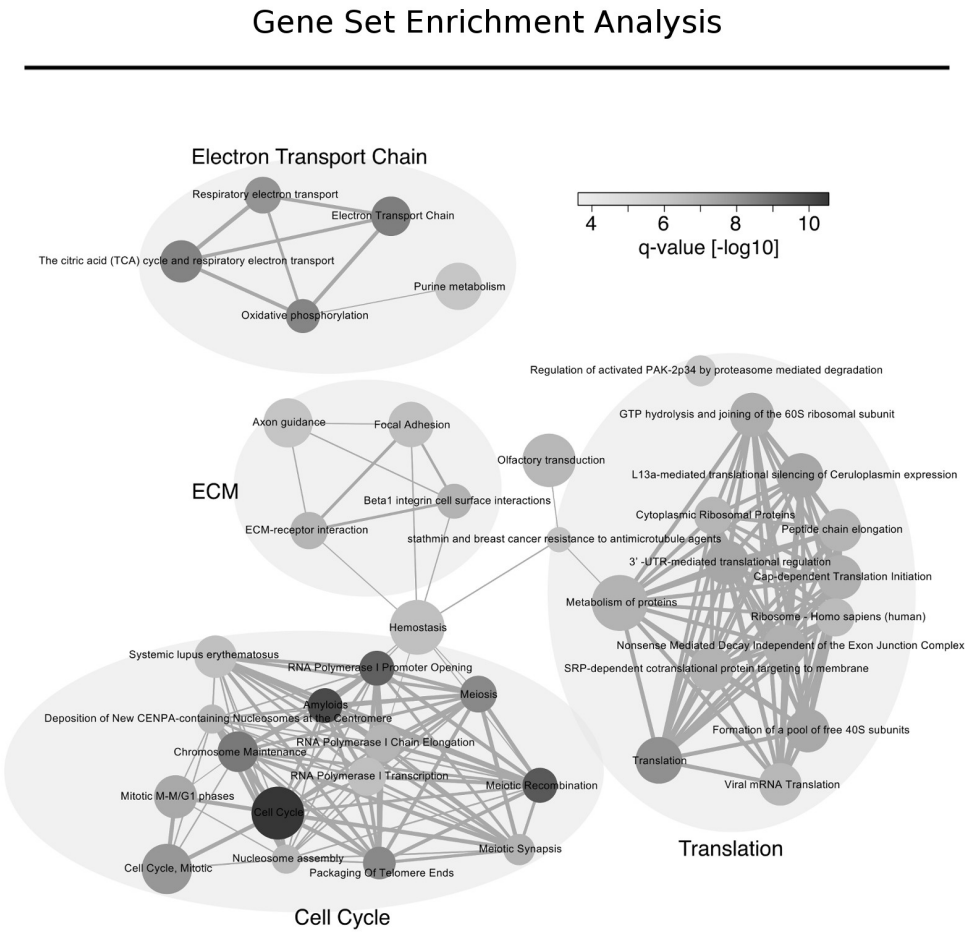


FIGURE 4.12: Gene set enrichment analysis of young and old fibroblast samples using gene sets from the Consensus Path DB (Kamburov et al., 2008). Shown are significantly down-regulated gene sets with age. Size and greyscale of the nodes are proportional to the gene set size and significance (Kalfalah et al., 2014).

4.4 Conclusion

The translation and interpretation of high-throughput measurements with respect to the categorical phenotypes commonly studied in the field of molecular and cell biology comes with various challenges. Both the GerontoSys and LungSys consortia have provided an opportunity to identify and examine the shortcomings in current workflows regarding the interpretation and translation of high-throughput data into scientific and medical insight. The observed shortcomings in the two case studies highlighted in the preceding sections include some which are difficult to address by computational approaches. Specifically, the challenge of limited sample sizes is dependent mostly on decisions with respect to the initial experimental design. In addition, the challenge of patient heterogeneity is not possible to be controlled by the experimental design. However, the prevalence of biological or technical bias that drives up the need for more samples is possible to be tackled. The aim of the BCN algorithm developed in Chapter 2 is to address this particular challenge. In addition, when communicating scientific results it is important to be able to provide an assessment of the biological relevance of candidate genes and phenotypical features identified by standard hypothesis tests. This is addressed by the MBR algorithm developed in Chapter 3. Thus, the case studies and workflows for the analysis of high-throughput experiments outlined in the preceding sections have identified important open challenges in quantitative biology and provide the motivation for the algorithms developed in this thesis.

Chapter 5

Outlook

In this thesis two algorithms have been presented that address the challenges of integrating large quantities of public high-throughput data. In quantitative biology, particularly in the field of molecular and cell biology, as well as medicine, such data is currently produced at a rapid rate (see Figure 1.5). The here developed algorithms are a step towards advancing high-throughput data integration by appropriate normalization and identification of biological relevance in hypothesis tests. However, several important open challenges exist that must be overcome in order for adoption of the proposed algorithms to occur and for high-throughput data integration to improve substantially.

5.1 Open challenges

Integration of sparse matrices

The blind compressive normalization algorithm proposed in Chapter 2 systematically corrects for technical and/or biological bias in the obtained high-throughput data to facilitate data integration. It does not require ad hoc assumptions about noise sources or biological signal common to unsupervised approaches. However, the scaling of the algorithm to large public databases is important for the absolute recovery performance to be optimal. While the scaling with respect to CPU time of the compressed sensing based solver routines is feasible for large high-dimensional databases, the memory consumption of these routines generally does not scale sufficiently. The current implementation of the blind compressive normalization algorithm is not able to exploit the advantages that come with sparse measurement operators. Specifically, the software dependencies for automatic differentiation do not allow for the integration of sparse matrices as data structures. Thus, the number of measurement operators and implicitly measurements is limited severely by the memory consumption of the compressed sensing based solver routines (Vandereycken, 2013). These routines are generally only scalable to a few hundred or maximally a few thousand samples (given an equal number of features) on an HPC system, due to the excessive memory consumption. Since measurement operators are typically 2-sparse in the proposed bias recovery approach, an ability to use sparse

matrices in the compressed sensing based solver routines would significantly impact the scalability and thus accuracy of the developed algorithm.

Integration of pursuit strategies

When using fixed rank constraints in matrix recovery problems, such as in the blind compressive normalization algorithm, two drawbacks exist. First, the fixed rank of the to be recovered low-rank matrix is generally not known *a priori*. Thus, recovery routines need to be run multiple times for different rank settings in order to determine the optimal rank *post hoc*. This is an inefficient approach and a significant drawback when contrasted to recovery methods based on nuclear norm regularization (Mazumder, Hastie, and Tibshirani, 2010). While the later methods are not flexible enough to be applied in the setting of blind recovery, an improvement in the efficiency of fixed rank methods is necessary to scale the developed blind compressive normalization algorithm to large databases on the order of hundredth of thousands to millions of features and samples. Secondly, inappropriate choices of the fixed rank parameter can result in ill-conditioned matrices for which compressed sensing based solver routines may converge slowly (Tan et al., 2014). To address these drawbacks, a pursuit type scheme has been developed recently. It is applicable to entry sensing as well as general recovery problems, such as those considered in the bias recovery of the developed blind compressive normalization algorithm (see Definition 1). The pursuit type scheme iteratively applies a nonlinear Riemannian conjugate gradient method to the recovery problem and converges under mild assumptions (Tan et al., 2014). From a theoretical perspective this approach can be understood as a warm start technique, much akin to those leveraged by nuclear norm regularization based approaches, such as softImpute (Mazumder, Hastie, and Tibshirani, 2010). Overall, implementation of a pursuit type scheme would improve the applicability of the developed blind compressive normalization algorithm significantly and further its adoption to a wider range of large scale blind recovery problems.

Integration of spike-in measurements

While appropriate quantitative standards are generally unavailable for high-throughput experiments in public databases, there has been an increase in the number of spike-in controls that are being measured. These incomplete standards can be important sources of additional information when it comes to the construction of absolute constraints in bias recovery. The developed blind compressive normalization algorithm leverages relative constraints in the form of dependencies. However, the addition of a certain number of absolute constraints obtained through spike-in controls may increase the performance and convergence speed of the algorithm significantly. An additional set of simulations may answer the question of how much such qualitatively different prior information improves the recovery performance and robustness of the developed algorithm.

Evaluation of cross-platform robustness

The measure of biological relevance developed in Chapter 3 is concerned with determining an appropriate null distribution for common test statistics applied to high-throughput data. Based on randomized public high-throughput data such null distributions can be precomputed for specific test statistics. The correction can give meaningful insights when one is interested in biological effects on the phenotypic level. However, an important factor influencing the null distribution is the measurement platform that the underlying data has been obtained from. Different approaches, such as those based on next generation sequencing or microarray technology, exhibit different biases with respect to specific features and measured intensity ranges. Thus, it remains to be investigated to what extend these factors influence the obtained measure of relevance. Such an investigation is important, if null distributions are precomputed on one technology and then subsequently applied to another. Also, it is reasonable to assume that a newly developed measurement technology does not provide sufficient public data yet to compute null distributions appropriately. Thus, pre-computed null distributions from other technologies may provide an alternative source of high-throughput data, if it can be shown through additional benchmarking simulations that such cross-platform application is sensible.

5.2 Next steps

Several of the open challenges highlighted in this section remain to be addressed in future works, especially for the proposed algorithms in Chapter 2 and Chapter 3 to be adopted and for high-throughput data integration to improve substantially. However, the additions of the necessary features currently limiting the blind normalization algorithm are not theoretical in nature and require essentially only an implementation and transfer of techniques already developed and tested in other areas of compressed sensing. Similarly, the outstanding analysis of cross-platform robustness of the here developed measure of biological relevance is possible to be completed without further theoretical work. Thus, with the annotated source code of the specific software packages made available (see Appendix A) it is possible to build upon the work presented in this thesis with the goal to meaningfully improve the integration of high-throughput databases in quantitative biology.

Appendix A

Supplementary material

A.1 Software packages

Blind Compressive Normalization (BCN)

The algorithm recovers bias from a high-throughput database without the use of prior information. Instead, detectable redundancies in the form of dependencies between samples and features are leveraged.

Source code

http://github.com/a378ec99/bcn

Dependencies

- scipy >= 0.19.0
- numpy >= 1.11.3
- scikit-learn >= 0.18.1
- pymanopt >= 0.2.3
- autograd >= 1.1.6
- scikit-image >= 0.13.1
- matplotlib >= 1.5.3 (optional – visualization)
- seaborn >= 0.7.1 (optional – visualization)
- mpi4py >= 3.0.0 (optional – parallel)

Licence

MIT License

Measure of Biological Relevance (MBR)

The algorithm rescales q-values of gene set enrichment tests based on null distributions computed from a high-throuhput database. This yields a measure of biological relevance for each test.

Source code

http://github.com/a378ec99/mbr

Dependencies

- scipy >= 0.19.0
- numpy >= 1.11.3
- h5py >= 2.6.0
- matplotlib >= 1.5.3
- seaborn >= 0.7.1

Licence

MIT License

A.2 Additional figures

Below are given additional figures that have been referenced in this thesis. These provide further details on the developed algorithms for the integration of high-throughput databases (Chapter 2 and Chapter 3) and the performed workflows of the discussed research consortia (Chapter 4).

[1]Wikimedia-Commons (2017), *Cross-validation Metrics*. Sourced on 21/08/17.
[2]Wikimedia-Commons (2016), *Principal Component Analysis*. Sourced on 23/08/17.

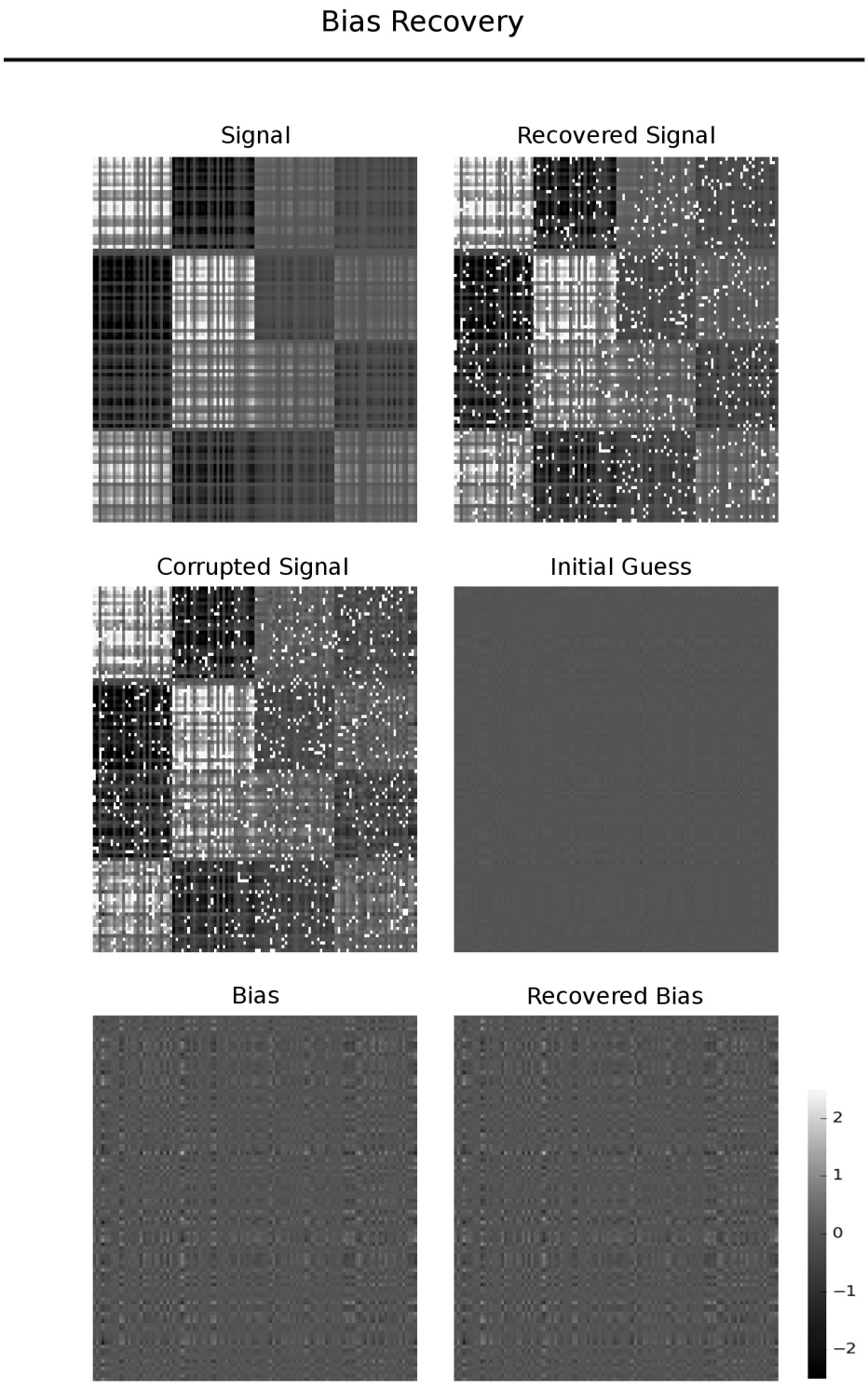


FIGURE A.1: The different stages of bias recovery with the BCN algorithm (see Chapter 2). Top left panel shows the true signal matched against the recovered signal on the right. The middle left panel shows the corrupted signal and the initial guess of the bias to be recovered on the right. Bottom left shows the true bias to be recovered matched with the recovered bias on the right. Blind recovery is highly accurate even though dependencies are unknown (missing values are not imputed).

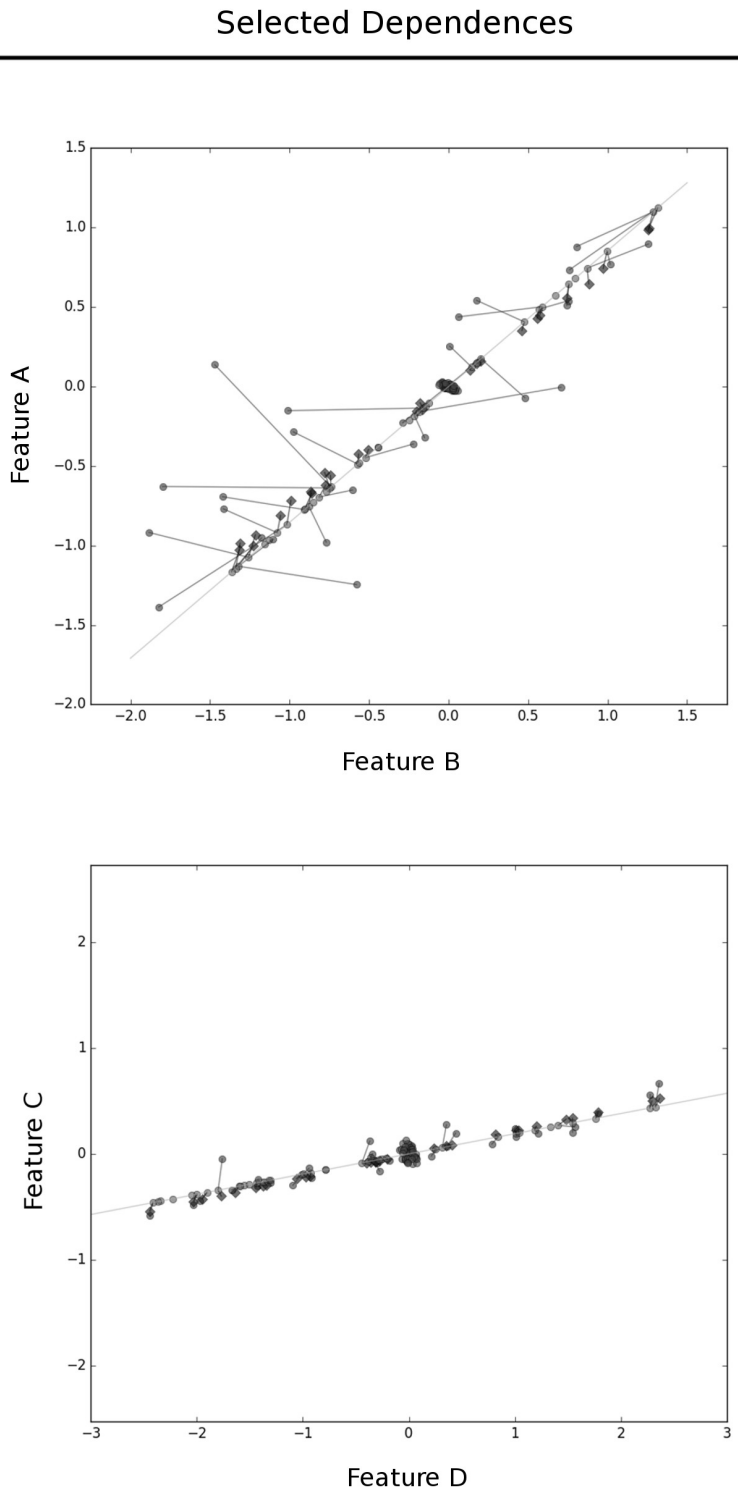


FIGURE A.2: Both the top and bottom panel show the successful bias recovery with the BCN algorithm in feature space of database rows A/B and C/D receptively (see Chapter 2). Squares denote the recovered signal, circles off of the diagonal the corrupted signal and circles on the diagonal the true signal. Connecting lines denote identical samples. Initial guesses of the solver are seen to be clustered around zero.

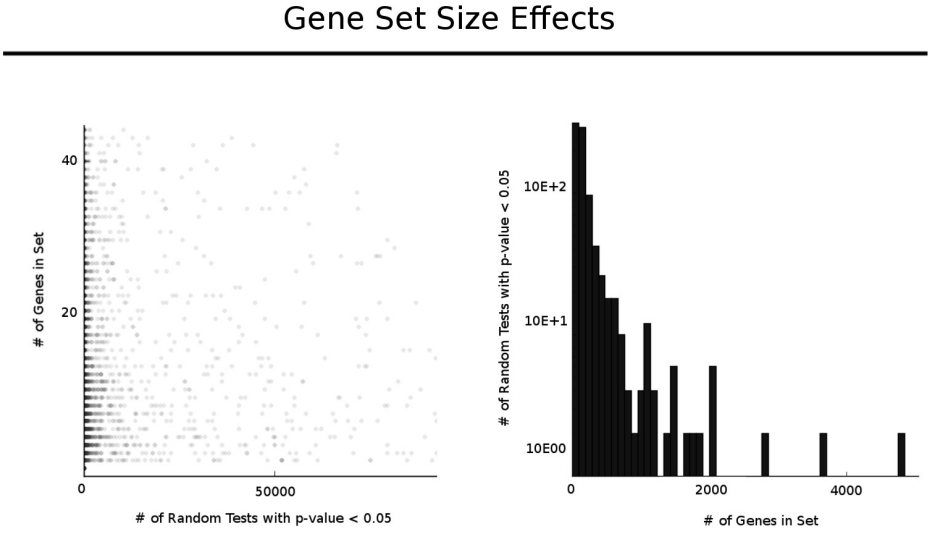


FIGURE A.3: Characterization of the effect of gene set size on the number of false positives identified with the MBR algorithm (see Chapter 3). On the right panel smaller gene sets appear to be more frequently identified as false positives. However, the proportion of small gene sets is also larger and thus overall no correlation between gene set size and false positive rate exists (left panel).

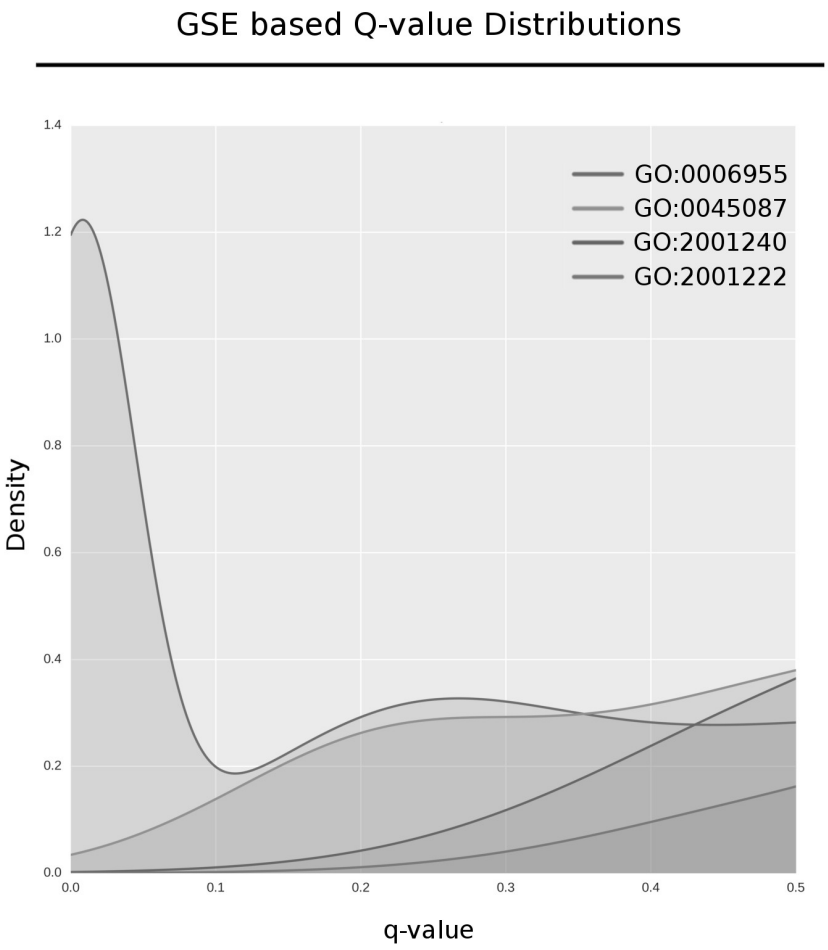


FIGURE A.4: Gene set enrichment based distributions of q-values for 4 distinct GO categories obtained from an analysis with the GAGE algorithm (Luo et al., 2009) across random contrasts of the GPL1261 high-throughput platform (Barrett et al., 2013). Notably, the distributions vary significantly between the GO categories shown, resulting in different correction factors with respect to biological relevance.

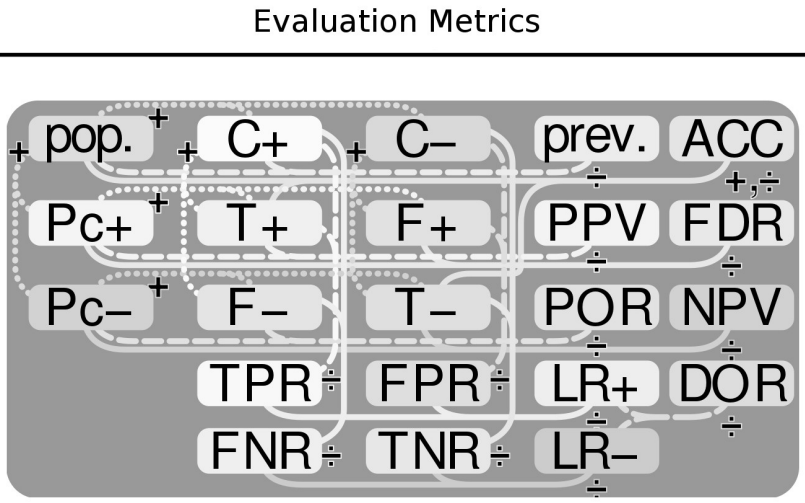


FIGURE A.5: Overview of cross-validation metrics and their relationship within a confusion matrix. See Section 4.2.3 for a detailed description of the given abbreviations. Modified with permission [1].

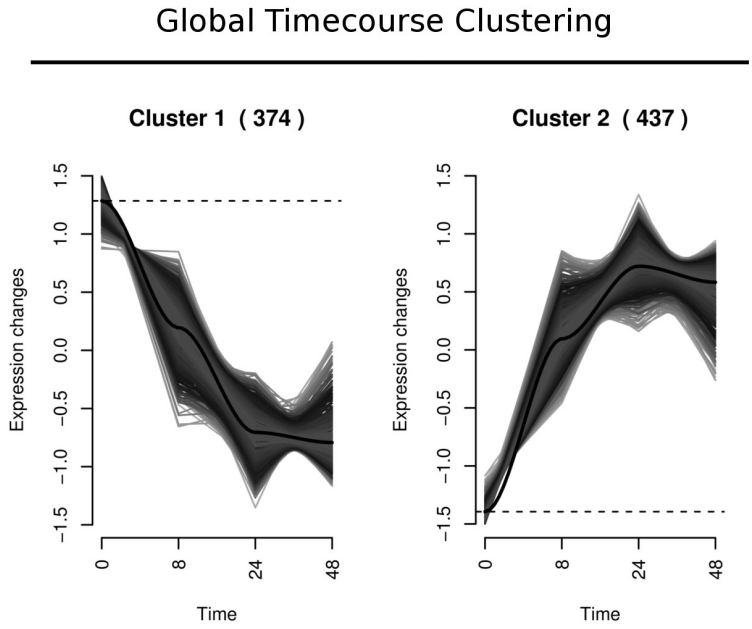


FIGURE A.6: Fuzzy noise robust soft-clustering (Futschik and Carlisle, 2005) of time-resolved RNA-Seq measurements of the transcriptome of TGFβ treated SK-MES1 cells. Two distinct clusters are visible over the course of 48 hours, one consisting of up-regulated genes (right) due to the TGFβ stimulus and the other of down-regulated genes (left). The Mfuzz software package was used with default parameters (Kumar and Futschik, 2007).

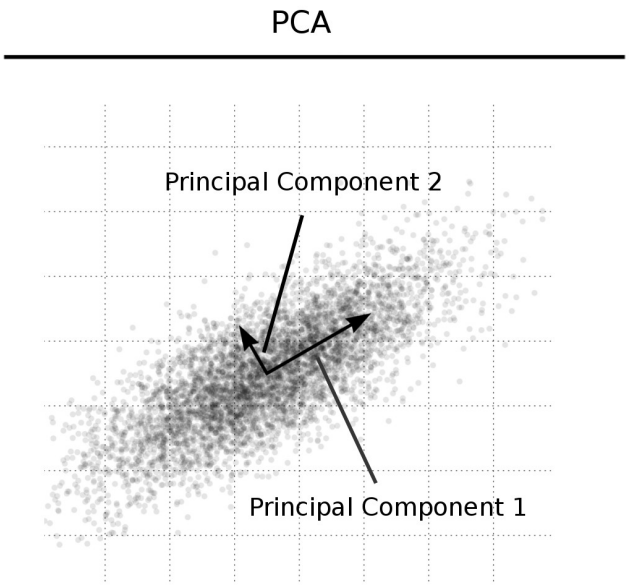
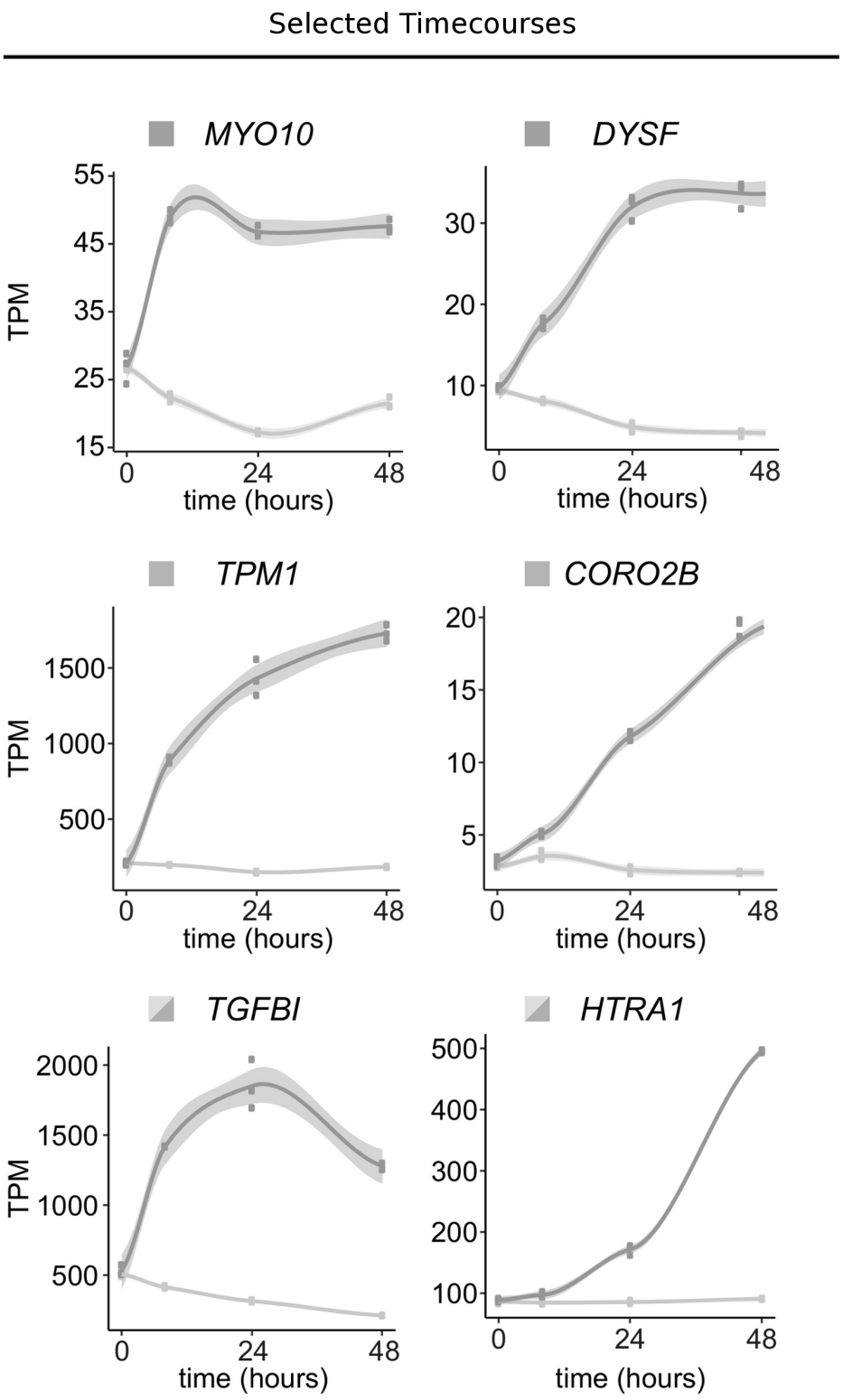
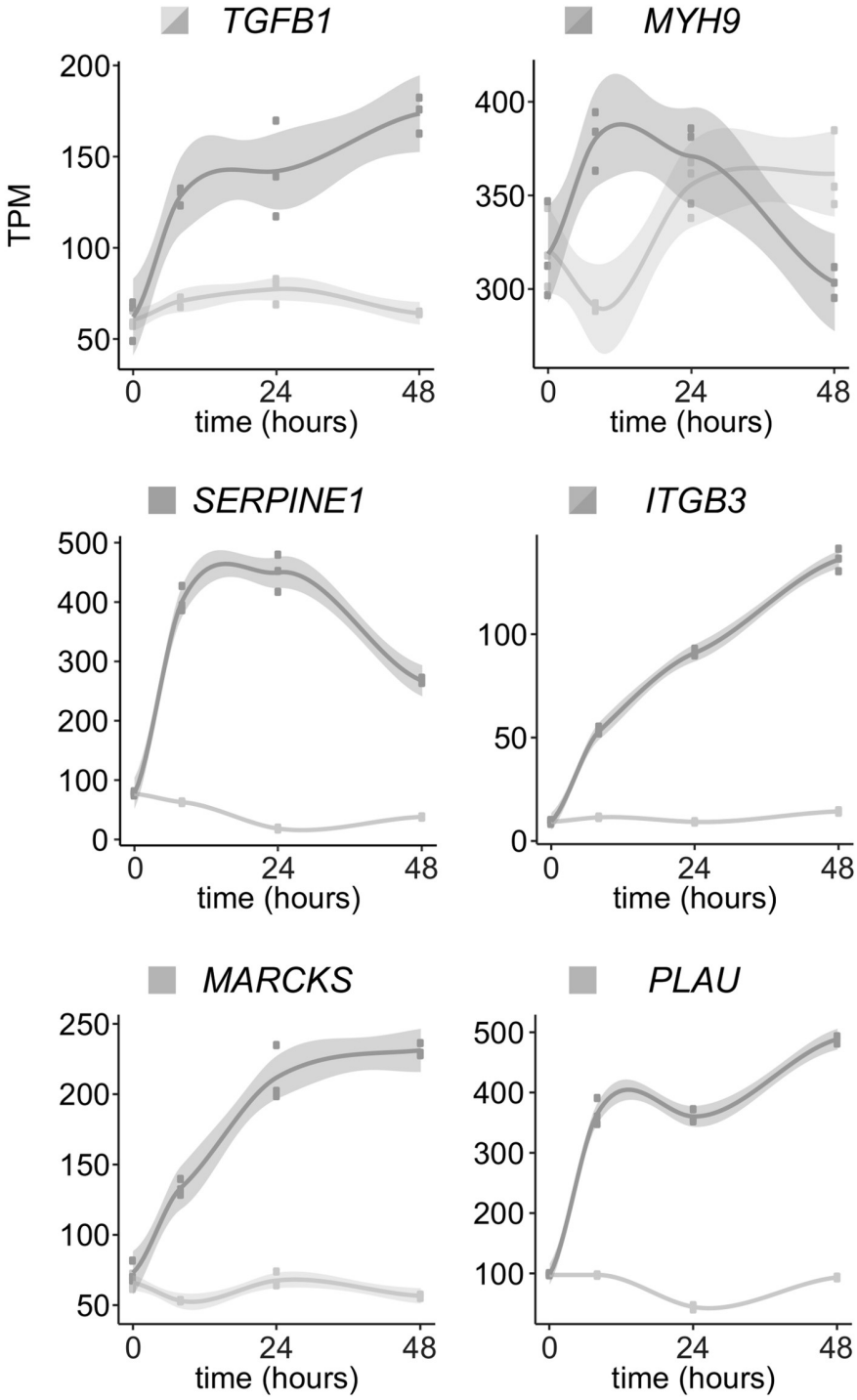


FIGURE A.7: Overview of principal component analysis (see Section 4.2.3) . The first principal component points in the direction of maximal variance. The subsequent components lie orthogonal to all previous components while also pointing into the direction of maximal variance within those constraints. The 2 dimensional dataset depicted here is denoted with features X and Y. In the case of high-dimensional data, typically Principal Component 1 and 2 are used as features to allow a reduction in dimensionality, while still visualizing most of the variance observed. Modified with permission [2].


Selected Timecourses
MYO10
DYSF
TPM1
CORO2B
TGFBI
HTRA1
TPM
time (hours)
0
24
48
55
45
35
25
15
30
20
10
1500
1000
500
2000
5
100
200
300
400


TGFB1
MYH9
SERPINE1
ITGB3
MARCKS
PLAU
TPM
time (hours)
0
24
48
50
100
150
200
300
350
400
500
250

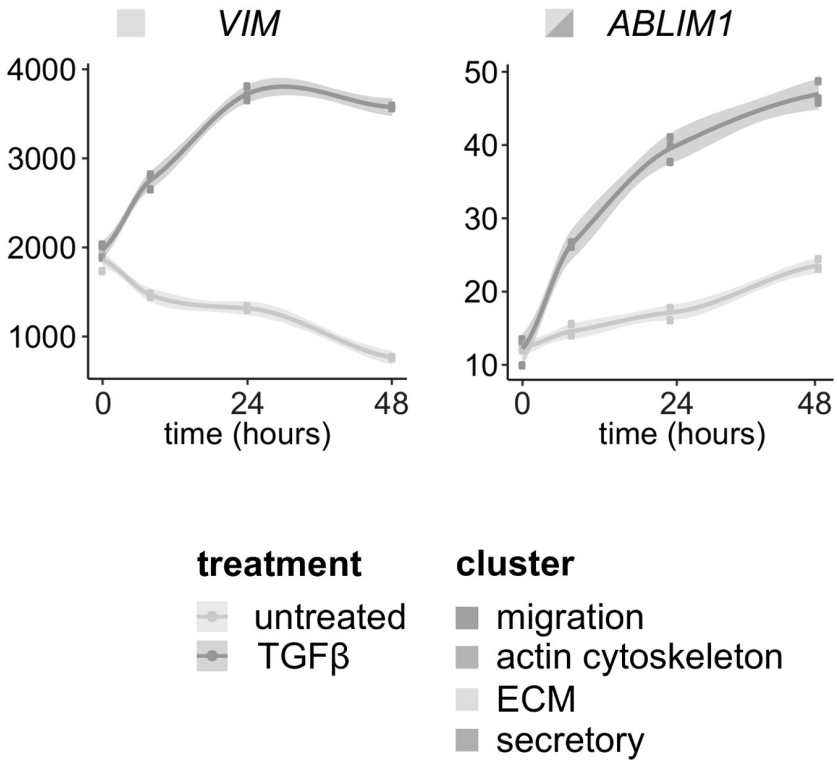


FIGURE A.8: Selected time-resolved RNA-Seq measurements of genes of interest upon TGFβ treatment in SK-MES1 cells. Cells were growth factor-depleted for three hours and stimulated with 2 ng/ml TGFβ1 or left untreated. The mRNA was extracted and sequenced using HiSeq 4000. Data is presented in TPM (transcripts per million) values. Each dot represents a biological replicate, shaded areas correspond to standard errors (C.1, Submitted).

Appendix B

Statement of contributions

Publications

A.1 I took the lead in the conceptualization and planning of the project, the theoretical algorithm development, as well as the implementation of the Python based software package (http://github.com/a378ec99/bcn). My contributions to the manuscript include the writing of the manuscript and creation of the contained figures, tables and supplementary material. Permission to reuse the figures for this thesis rest with me until the manuscript is accepted. Parts of this work was presented by me at the LungSys conferences (2014, 2015, 2016), the Herzogenhorn retreat of Prof. Rolf Backofen (2014, 2016) and the IMMZ institute retreat (2014).

A.2 I was involved in conceptualizing and planning of the project, conducting a literature review and presenting its outcome to my co-authors in order to motivate research on this topic. I contributed to the manuscript section 3 and figure 3. I was significantly involved in the overall proofing of the manuscript. The majority of this work was conducted under the research stipend "Forschungsstipendien für Systembiologen" (BMBF #0316042G). Permissions to reuse the figures for this thesis are based on the Copyright Clearance Center (Account # 3001216592). The findings of this research were presented by me at the PRBB seminar series of Prof. Jordi Ojalvo-Garcia and Dr. Lucas Carey (2014).

B.1 I conducted the high-throughput data analysis for the project, including the preliminary gene expression analysis, normalization of the high-throughput data, analysis with LIMMA for the statistical identification of candidate genes, gene set enrichment tests and principal component analyses. In addition, I managed the database storage and maintenance of clinical experiments, assured the quality of annotations and corresponded with members of the LungSys consortium regarding biostatistical questions. To the manuscript I contributed the materials and methods section and figure 1. Permissions to reuse the figures for this thesis are based on the Copyright Clearance Center (Account # 3001216592).

B.2 I conducted the high-throughput data analysis for the project, including the preliminary gene expression analysis, normalization of the high-throughput data, analysis with LIMMA for the statistical identification of candidate genes, gene set enrichment tests and principal component analyses. In addition, I managed the database storage and maintenance of clinical experiments, assured the quality of annotations and corresponded with members of the LungSys consortium regarding biostatistical questions. To the manuscript I contributed the materials and methods section and figures 2, 3 and 5. Permissions to reuse the figures for this thesis are based on the Copyright Clearance Center (Account # 3001216592).

C.1 I was involved in the planning, experimental design and analysis with respect to the high-throughput measurements obtained. To the manuscript I contributed figure 3, the analytical methods section and several proof readings as well as modifications of the manuscript. Throughout the project I was leading biostatistical analyses and conducted preliminary and final gene expression analyses, normalization of the high-throughput data, LIMMA analyses for the statistical identification of candidate genes, fuzzy clustering of gene expression time series, gene set enrichment tests and consortium correspondence. The work was presented by me at the LungSys conferences (2014, 2015, 2016). Permission to reuse the respective figures for this thesis were obtained from Dr. Dmytro Dvornikov.

C.2 I conducted the high-throughput data analysis for the project, including the preliminary gene expression analysis, normalization of the high-throughput data, analysis with LIMMA for the statistical identification of candidate genes, gene set enrichment tests and principal component analyses. To the manuscript I contributed figure 6A and details of the methods section.

Bibliography

Aach, John and George M Church (2001). "Aligning gene expression time series with time warping algorithms". In: *Bioinformatics* 17.6, pp. 495–508.

Affymetrix, I (2007). "GeneChip Gene 1.0 ST Array System for Human, Mouse and Rat. A simple and affordable solution for advanced gene-level expression profiling. Data sheet". In: *Affymetrix Data Sheet*.

Ahmed, Ali, Benjamin Recht, and Justin Romberg (2014). "Blind deconvolution using convex programming". In: *IEEE Transactions on Information Theory* 60.3, pp. 1711–1732.

Aickin, Mikel and Helen Gensler (1996). "Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods". In: *American journal of public health* 86.5, pp. 726–728.

Akaike, Hirotugu (1974). "A new look at the statistical model identification". In: *IEEE transactions on automatic control* 19.6, pp. 716–723.

Aldana, Maximino et al. (2007). "Robustness and evolvability in genetic regulatory networks". In: *Journal of theoretical biology* 245.3, pp. 433–448.

Allen, Genevera I and Robert Tibshirani (2012). "Inference with transposable data: modelling the effects of row and column correlations". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.4, pp. 721–743.

Arumugam, Thiruvengadam et al. (2009). "Epithelial to mesenchymal transition contributes to drug resistance in pancreatic cancer". In: *Cancer research* 69.14, pp. 5820–5828.

Ashburner, Michael et al. (2000). "Gene Ontology: tool for the unification of biology". In: *Nature genetics* 25.1, pp. 25–29.

Balleza, Enrique et al. (2008). "Critical dynamics in genetic regulatory networks: examples from four kingdoms". In: *PLoS One* 3.6, e2456.

Barabasi, Albert-Laszlo and Zoltan N Oltvai (2004). "Network biology: understanding the cell's functional organization". In: *Nature reviews. Genetics* 5.2, p. 101.

Barnes, Michael et al. (2005). "Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms". In: *Nucleic acids research* 33.18, pp. 5914–5923.

Barrett, Tanya et al. (2013). "NCBI GEO: archive for functional genomics data sets – update". In: *Nucleic acids research* 41.D1, pp. D991–D995.

Beck, Amir and Marc Teboulle (2009). "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". In: *SIAM journal on imaging sciences* 2.1, pp. 183–202.

Benito, Monica et al. (2004). "Adjustment of systematic microarray data biases". In: *Bioinformatics* 20.1, pp. 105–114.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300.

Berinde, Radu and Piotr Indyk (2008). "Sparse recovery using sparse random matrices". In: *preprint*.

Berridge, Michael J, Martin D Bootman, and H Llewelyn Roderick (2003). "Calcium signalling: dynamics, homeostasis and remodelling". In: *Nature reviews. Molecular cell biology* 4.7, p. 517.

Binder, Hans and Stephan Preibisch (2008). "Hook-calibration of GeneChip microarrays: Theory and algorithm". In: *Algorithms for Molecular Biology* 3.1, p. 1.

Blanchard, Jeffrey D, Jared Tanner, and Ke Wei (2015). "CGIHT: conjugate gradient iterative hard thresholding for compressed sensing and matrix completion". In: *Information and Inference: A Journal of the IMA* 4.4, pp. 289–327.

Blumensath, Thomas and Mike E Davies (2009). "Iterative hard thresholding for compressed sensing". In: *Applied and computational harmonic analysis* 27.3, pp. 265–274.

— (2010). "Normalized iterative hard thresholding: Guaranteed stability and performance". In: *IEEE Journal of selected topics in signal processing* 4.2, pp. 298–309.

Bolstad, Benjamin M et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias". In: *Bioinformatics* 19.2, pp. 185–193.

Bornholdt, Stefan and Thimo Rohlf (2000). "Topological evolution of dynamical networks: Global criticality from local dynamics". In: *Physical Review Letters* 84.26, p. 6114.

Cai, Jian-Feng, Emmanuel J Candès, and Zuowei Shen (2010). "A singular value thresholding algorithm for matrix completion". In: *SIAM Journal on Optimization* 20.4, pp. 1956–1982.

Cai, T Tony, Anru Zhang, et al. (2015). "ROP: Matrix recovery via rank-one projections". In: *The Annals of Statistics* 43.1, pp. 102–138.

Candes, Emmanuel and Justin Romberg (2007). "Sparsity and incoherence in compressive sampling". In: *Inverse problems* 23.3, p. 969.

Candes, Emmanuel J and Yaniv Plan (2010). "Matrix completion with noise". In: *Proceedings of the IEEE* 98.6, pp. 925–936.

— (2011). "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements". In: *IEEE Transactions on Information Theory* 57.4, pp. 2342–2359.

Candès, Emmanuel J and Michael B Wakin (2008). "An introduction to compressive sampling". In: *IEEE signal processing magazine* 25.2, pp. 21–30.

Cheadle, Chris et al. (2003). "Analysis of microarray data using Z score transformation". In: *The Journal of molecular diagnostics* 5.2, pp. 73–81.

Chen, CH et al. (2017). "Upregulation of MARCKS in kidney cancer and its potential as a therapeutic target". In: *Oncogene*.

Chen, Scott Shaobing, David L Donoho, and Michael A Saunders (2001). "Atomic decomposition by basis pursuit". In: *SIAM review* 43.1, pp. 129–159.

Coifman, Ronald, F Geshwind, and Yves Meyer (2001). "Noiselets". In: *Applied and Computational Harmonic Analysis* 10.1, pp. 27–44.

Consortium, Gene Ontology et al. (2015). "Gene ontology consortium: going forward". In: *Nucleic acids research* 43.D1, pp. D1049–D1056.

Crick, Francis (1970). "Central dogma of molecular biology". In: *Nature* 227.5258, pp. 561–563.

Cristianini, Nello and John Shawe-Taylor (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Croft, David et al. (2014). "The Reactome pathway knowledgebase". In: *Nucleic acids research* 42.D1, pp. D472–D477.

Davenport, Mark A et al. (2007). "The smashed filter for compressive classification and target recognition". In: *Computational Imaging V at SPIE Electronic Imaging*.

Do, Chuong B and Serafim Batzoglou (2008). "What is the expectation maximization algorithm?" In: *Nature biotechnology* 26.8, pp. 897–899.

Doms, Andreas and Michael Schroeder (2005). "GoPubMed: exploring PubMed with the gene ontology". In: *Nucleic acids research* 33.suppl_2, W783–W786.

Donoho, David L et al. (2012). "Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit". In: *IEEE Transactions on Information Theory* 58.2, pp. 1094–1121.

Drilon, Alexander et al. (2012). "Squamous-cell carcinomas of the lung: emerging biology, controversies, and the promise of targeted therapy". In: *The lancet oncology* 13.10, e418–e426.

Drossel, Barbara (2008). "Random boolean networks". In: *Reviews of nonlinear dynamics and complexity* 1, pp. 69–110.

Duarte, Marco F et al. (2008). "Single-pixel imaging via compressive sampling". In: *IEEE signal processing magazine* 25.2, pp. 83–91.

EFSA (2011). "Statistical significance and biological relevance". In: *EFSA Journal* 9.9.

Ester, Martin et al. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Kdd*. Vol. 96. 34, pp. 226–231.

Fawcett, Tom (2006). "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8, pp. 861–874.

Fox, Jeffrey J and Colin C Hill (2001). "From topology to dynamics in biochemical networks". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 11.4, pp. 809–815.

Frane, Andrew V (2015). "Are per-family type I error rates relevant in social and behavioral science?" In: *Journal of Modern Applied Statistical Methods* 14.1, p. 5.

Frey, Brendan J and Delbert Dueck (2007). "Clustering by passing messages between data points". In: *science* 315.5814, pp. 972–976.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. Springer series in statistics New York.

Fujisawa, Katsuki et al. (2000). "Numerical evaluation of SDPA (semidefinite programming algorithm)". In: *High performance optimization*. Springer, pp. 267–301.

Futschik, Matthias E and Bronwyn Carlisle (2005). "Noise-robust soft clustering of gene expression time-course data". In: *Journal of bioinformatics and computational biology* 3.04, pp. 965–988.

Garibay, Gorka Ruiz de et al. (2015). "Lymphangioleiomyomatosis biomarkers linked to lung metastatic potential and cell stemness". In: *PloS one* 10.7, e0132546.

GerontoSys Consortium (2017). *Systembiologie für die Gesundheit im Alter – GerontoSys*. URL: http://www.ptj.de/gerontosys (visited on 07/28/2017).

Gershenson, Carlos (2012). "Guiding the self-organization of random Boolean networks". In: *Theory in Biosciences* 131.3, pp. 181–191.

Ghosh, Malay, David M Nickerson, and Pranab K Sen (1987). "Sequential shrinkage estimation". In: *The Annals of Statistics*, pp. 817–829.

GO Consortium (2004). "The Gene Ontology (GO) database and informatics resource". In: *Nucleic acids research* 32.suppl 1, pp. D258–D261.

Goldberg, Aaron D, C David Allis, and Emily Bernstein (2007). "Epigenetics: a landscape takes shape". In: *Cell* 128.4, pp. 635–638.

Haider, Syed et al. (2014). "A multi-gene signature predicts outcome in patients with pancreatic ductal adenocarcinoma". In: *Genome medicine* 6.12, p. 105.

Hallett, Robin M et al. (2012). "A gene signature for predicting outcome in patients with basal-like breast cancer". In: *Scientific reports* 2, p. 227.

Han, Jiawei, Jian Pei, and Micheline Kamber (2011). *Data mining: concepts and techniques*. Elsevier.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC Press.

Hastie, Trevor et al. (2015). "Matrix completion and low-rank SVD via fast alternating least squares". In: *Journal of Machine Learning Research* 16, pp. 3367–3402.

Haug, Kenneth et al. (2012). "MetaboLights – an open-access general-purpose repository for metabolomics studies and associated meta-data". In: *Nucleic acids research*, gks1004.

Hecker (2009). "Gene regulatory network inference: Data integration in dynamic models – A review". In: *Biosystems*.

Hesse, Janina and Thilo Gross (2014). "Self-organized criticality as a fundamental property of neural systems". In: *Frontiers in systems neuroscience* 8.

Hornung, Roman, Anne-Laure Boulesteix, and David Causeur (2016). "Combining location-and-scale batch effect adjustment with data cleaning by latent factor adjustment". In: *BMC bioinformatics* 17.1, p. 1.

Houle, David et al. (2011). "Measurement and meaning in biology". In: *The Quarterly Review of Biology* 86.1, pp. 3–34.

Huber, Wolfgang et al. (2002). "Variance stabilization applied to microarray data calibration and to the quantification of differential expression". In: *Bioinformatics* 18.suppl 1, S96–S104.

iGEM (2017). *Standard Registry of Biological Parts*. URL: http://parts.igem.org.

Irizarry, Rafael A et al. (2003). "Summaries of Affymetrix GeneChip probe level data". In: *Nucleic acids research* 31.4, e15–e15.

Jain, Prateek, Praneeth Netrapalli, and Sujay Sanghavi (2013). "Low-rank matrix completion using alternating minimization". In: *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM, pp. 665–674.

Johnson, W Evan, Cheng Li, and Ariel Rabinovic (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1, pp. 118–127.

Joyce, Andrew R and Bernhard Ø Palsson (2006). "The model organism as a system: integrating'omics' data sets". In: *Nature Reviews Molecular Cell Biology* 7.3, pp. 198–210.

Kalaitzis, Alfredo A and Neil D Lawrence (2011). "A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression". In: *BMC bioinformatics* 12.1, p. 180.

Kalfalah, Faiza et al. (2014). "Inadequate mito-biogenesis in primary dermal fibroblasts from old humans is associated with impairment of PGC1A-independent stimulation". In: *Experimental gerontology* 56, pp. 59–68.

Kalfalah, Faiza et al. (2015). "Structural chromosome abnormalities, increased DNA strand breaks and DNA strand break repair deficiency in dermal fibroblasts from old female human donors". In: *Aging (Albany NY)* 7.2, p. 110.

Kamburov, Atanas et al. (2008). "ConsensusPathDB—a database for integrating human functional interaction networks". In: *Nucleic acids research* 37.suppl_1, pp. D623–D628.

Karr, Jonathan R et al. (2012). "A whole-cell computational model predicts phenotype from genotype". In: *Cell* 150.2, pp. 389–401.

Katono, Ken et al. (2015). "Prognostic significance of MYH9 expression in resected non-small cell lung cancer". In: *PloS one* 10.3, e0121460.

Kauffman, Stuart A (1969). "Metabolic stability and epigenesis in randomly constructed genetic nets". In: *Journal of theoretical biology* 22.3, pp. 437–467.

Kaufman, Leonard and Peter J Rousseeuw (2009). *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons.

Kim, Edward S et al. (2008). "Gefitinib versus docetaxel in previously treated non-small-cell lung cancer (INTEREST): a randomised phase III trial". In: *The Lancet* 372.9652, pp. 1809–1818.

Kirk, Paul DW and Michael PH Stumpf (2009). "Gaussian process regression bootstrapping: exploring the effects of uncertainty in time course data". In: *Bioinformatics* 25.10, pp. 1300–1306.

Krantz, David H et al. (1971). *Foundations of Measurement (Additive and Polynomial Representations), vol. 1*.

Kumar, Lokesh and Matthias E Futschik (2007). "Mfuzz: a software package for soft clustering of microarray data". In: *Bioinformation* 2.1, p. 5.

Lamouille, Samy, Jian Xu, and Rik Derynck (2014). "Molecular mechanisms of epithelial mesenchymal transition". In: *Nature reviews Molecular cell biology* 15.3, pp. 178–196.

Lauden, Laura et al. (2014). "TGF-β-induced (TGFBI) protein in melanoma: a signature of high metastatic potential". In: *Journal of Investigative Dermatology* 134.6, pp. 1675–1685.

Lazar, Cosmin et al. (2012). "Batch effect removal methods for microarray gene expression data integration: a survey". In: *Briefings in bioinformatics*, bbs037.

Leek, Jeffrey T et al. (2012). "The sva package for removing batch effects and other unwanted variation in high-throughput experiments". In: *Bioinformatics* 28.6, pp. 882–883.

Legenstein, Robert and Wolfgang Maass (2007). "What makes a dynamical system computationally powerful". In: *New directions in statistical signal processing: From systems to brain*, pp. 127–154.

Li, Cheng and Wing Hung Wong (2001). "Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application". In: *Genome biology* 2.8, p. 1.

Li, Ping, Trevor J Hastie, and Kenneth W Church (2006). "Very sparse random projections". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 287–296.

Li, Ping and Cun-Hui Zhang (2015). "Compressed sensing with very sparse gaussian random projections". In: *Artificial Intelligence and Statistics*, pp. 617–625.

Liaw, Andy, Matthew Wiener, et al. (2002). "Classification and regression by randomForest". In: *R news* 2.3, pp. 18–22.

Ling, Shuyang and Thomas Strohmer (2015). "Self-calibration and biconvex compressive sensing". In: *Inverse Problems* 31.11, p. 115002.

Link, Hannes et al. (2015). "Real-time metabolome profiling of the metabolic switch between starvation and growth". In: *nAture methods* 12.11, pp. 1091–1097.

Lizier, Joseph T, Siddharth Pritam, and Mikhail Prokopenko (2011). "Computational capabilities of small-world Boolean networks". In: *Advances in Artificial Life, ECAL*, pp. 463–464.

Look, Maxime P et al. (2002). "Pooled analysis of prognostic impact of urokinase-type plasminogen activator and its inhibitor PAI-1 in 8377 breast cancer patients". In: *Journal of the National Cancer Institute* 94.2, pp. 116–128.

Lovell, David P (2013). "Biological importance and statistical significance". In: *Journal of agricultural and food chemistry* 61.35, pp. 8340–8348.

Lovén, Jakob et al. (2012). "Revisiting global gene expression analysis". In: *Cell* 151.3, pp. 476–482.

LungSys Consortium (2017). *Systems Biology - A novel approach to The Lung Cancer Problem*. URL: http://www.lungsys.de (visited on 07/28/2017).

Luo, Weijun et al. (2009). "GAGE: generally applicable gene set enrichment for pathway analysis". In: *BMC bioinformatics* 10.1, p. 161.

Luque, B and RV Solé (1997). "Controlling chaos in random Boolean networks". In: *EPL (Europhysics Letters)* 37.9, p. 597.

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9.Nov, pp. 2579–2605.

Maeng, Young-In et al. (2014). "Transcription factors related to epithelial mesenchymal transition in tumor center and margin in invasive lung adenocarcinoma". In: *International journal of clinical and experimental pathology* 7.7, p. 4095.

Maier, Tobias, Marc Güell, and Luis Serrano (2009). "Correlation of mRNA and protein in complex biological samples". In: *FEBS letters* 583.24, pp. 3966–3973.

Makowska, Katarzyna A et al. (2015). "Specific myosins control actin organization, cell morphology, and migration in prostate cancer cells". In: *Cell reports* 13.10, pp. 2118–2125.

Mallat, Stéphane and Zhifeng Zhang (1993). *Matching pursuit with time-frequency dictionaries*. Tech. rep. Courant Institute of Mathematical Sciences New York United States.

MAQC Consortium (2006). "The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements". In: *Nature biotechnology* 24.9, p. 1151.

Marwitz, Sebastian et al. (2016). "Downregulation of the TGFβ Pseudoreceptor BAMBI in Non–Small Cell Lung Cancer Enhances TGFβ Signaling and Invasion". In: *Cancer research* 76.13, pp. 3785–3801.

Mazumder, Rahul, Trevor Hastie, and Robert Tibshirani (2010). "Spectral regularization algorithms for learning large incomplete matrices". In: *Journal of machine learning research* 11.Aug, pp. 2287–2322.

McCall, Matthew N., Benjamin M. Bolstad, and Rafael A Irizarry (2010). "Frozen robust multiarray analysis (fRMA)". en. In: *Biostatistics* 11.2, pp. 242–253. ISSN: 1465-4644, 1468-4357. (Visited on 11/30/2015).

McCall, Matthew N et al. (2011). "The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes". In: *Nucleic acids research* 39.suppl 1, pp. D1011–D1015.

Mika, Sebastian et al. (1999). "Kernel PCA and de-noising in feature spaces". In: *Advances in neural information processing systems*, pp. 536–542.

Moffitt, Richard A et al. (2011). "caCORRECT2: Improving the accuracy and reliability of microarray data in the presence of artifacts". In: *BMC bioinformatics* 12.1, p. 1.

Montgomery, Douglas C (2008). *Design and analysis of experiments*. John Wiley & Sons.

Munkres, James R (2000). *Topology*. Prentice Hall.

Naldini, Luigi (2015). "Gene therapy returns to centre stage". In: *Nature* 526.7573, p. 351.

Needell, Deanna and Joel A Tropp (2009). "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples". In: *Applied and computational harmonic analysis* 26.3, pp. 301–321.

Nykter, Matti et al. (2008). "Gene expression dynamics in the macrophage exhibit criticality". In: *Proceedings of the National Academy of Sciences* 105.6, pp. 1897–1900.

Ohse S, Dvornikov D Schneider MA Szczygieł M Titkova I Rosenblatt M Muley T Warth A Herth FJ Dienemann H Thomas M Timmer J Schilling M Busch H Boerries M Meister M & Klingmüller U (2018). "Expression ratio of the TGFβ-inducible gene MYO10 is prognostic for overall survival of squamous cell lung cancer patients and predicts chemotherapy response". In: *Scientific Reports* 8, pp. 9517–9530.

Ohse, S, M Börries, and H Busch (2019). "Blind normalization of public high-throughput databases". In: *PeerJ Computer Science* 5, pp. 231–247.

Ouderkirk, Jessica L and Mira Krendel (2014). "Non-muscle myosins in tumor progression, cancer cell invasion, and metastasis". In: *Cytoskeleton* 71.8, pp. 447–463.

Parker, Hilary S, Héctor Corrada Bravo, and Jeffrey T Leek (2014). "Removing batch effects for prediction problems with frozen surrogate variable analysis". In: *PeerJ* 2, e561.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* 12.Oct, pp. 2825–2830.

Pepelyshev, Andrey (2010). "The role of the nugget term in the Gaussian process method". In: *mODa 9–Advances in Model-Oriented Design and Analysis*. Springer, pp. 149–156.

Piccolo, Stephen R. et al. (2012). "A single-sample microarray normalization method to facilitate personalized-medicine workflows". eng. In: *Genomics* 100.6, pp. 337–344. ISSN: 1089-8646.

Pilastri, André Luiz and Joao Manuel RS Tavares (2016). "Reconstruction Algorithms in Compressive Sensing: An Overview". In: *11th edition of the Doctoral Symposium in Informatics Engineering (DSIE¿ 16)*.

Poblanno-Balp, Rodrigo and Carlos Gershenson (2011). "Modular random boolean networks1". In: *Artificial life* 17.4, pp. 331–351.

Pomerance, Andrew et al. (2009). "The effect of network topology on the stability of discrete state models of genetic control". In: *Proceedings of the National Academy of Sciences* 106.20, pp. 8209–8214.

Potra, Florian A and Stephen J Wright (2000). "Interior-point methods". In: *Journal of Computational and Applied Mathematics* 124.1-2, pp. 281–302.

Powers, David Martin (2011). "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". In:

Press, William H (2007). *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press.

Rasmussen, Carl Edward and Christopher KI Williams (2006). *Gaussian processes for machine learning*. Vol. 1. MIT press Cambridge.

Recht, Benjamin, Maryam Fazel, and Pablo A Parrilo (2010). "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization". In: *SIAM review* 52.3, pp. 471–501.

Reuter, Jason A, Damek V Spacek, and Michael P Snyder (2015). "High-throughput sequencing technologies". In: *Molecular cell* 58.4, pp. 586–597.

Ritchie, Matthew E et al. (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic acids research* 43.7, e47–e47.

Rivasplata, Omar (2012). "Subgaussian random variables: An expository note". In: *Internet publication*.

Rooney, Melissa, Siddhartha Devarakonda, and Ramaswamy Govindan (2013). "Genomics of squamous cell lung cancer". In: *The oncologist* 18.6, pp. 707–716.

Sandberg, Rickard and Ola Larsson (2007). "Improved precision and accuracy for microarrays using updated probe set definitions". In: *BMC bioinformatics* 8.1, p. 48.

Schaffter, Thomas, Daniel Marbach, and Dario Floreano (2011). "GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods". In: *Bioinformatics* 27.16, pp. 2263–2270.

Seita, Jun et al. (2012). "Gene Expression Commons: an open platform for absolute gene expression profiling". In: *PloS one* 7.7, e40321.

Serra, Roberto, Marco Villani, and Alessandro Semeria (2004). "Genetic network models and statistical properties of gene expression data in knock-out experiments". In: *Journal of theoretical biology* 227.1, pp. 149–157.

Shabalin, Andrey A et al. (2008). "Merging two gene-expression studies via cross-platform normalization". In: *Bioinformatics* 24.9, pp. 1154–1160.

Shaffer, Juliet Popper (1995). "Multiple hypothesis testing". In: *Annual review of psychology* 46.1, pp. 561–584.

Shintani, Yasushi et al. (2011). "Epithelial to mesenchymal transition is a determinant of sensitivity to chemoradiotherapy in non-small cell lung cancer". In: *The Annals of thoracic surgery* 92.5, pp. 1794–1804.

Shmulevich, Ilya and Stuart A Kauffman (2004). "Activities and sensitivities in Boolean network models". In: *Physical review letters* 93.4, p. 048701.

Shmulevich, Ilya, Stuart A Kauffman, and Maximino Aldana (2005). "Eukaryotic cells are dynamically ordered or critical but not chaotic". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.38, pp. 13439–13444.

Singh, Sheila K et al. (2003). "Identification of a cancer stem cell in human brain tumors". In: *Cancer research* 63.18, pp. 5821–5828.

Smyth, Gordon K (2005). "Limma: linear models for microarray data". In: *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, pp. 397–420.

Solé, Ricard V and Sergi Valverde (2006). "Are network motifs the spandrels of cellular complexity?" In: *Trends in ecology & evolution* 21.8, pp. 419–422.

— (2008). "Spontaneous emergence of modularity in cellular networks". In: *Journal of The Royal Society Interface* 5.18, pp. 129–133.

Soltermann, Alex et al. (2008). "Prognostic significance of epithelial mesenchymal and mesenchymal epithelial transition protein expression in non–small cell lung cancer". In: *Clinical Cancer Research* 14.22, pp. 7430–7437.

Sterlacci, William et al. (2012). "High transforming growth factor β expression represents an important prognostic parameter for surgically resected non–small cell lung cancer". In: *Human pathology* 43.3, pp. 339–349.

Stevens, SS (1946). *On the Theory of Scales of Measurement*.

Stöger, Dominik, Peter Jung, and Felix Krahmer (2016). "Blind deconvolution and compressed sensing". In: *Compressed Sensing Theory and its Applications to Radar, Sonar and Remote Sensing (CoSeRa), 2016 4th International Workshop on*. IEEE, pp. 24–27.

Sturm, Jos F (1999). "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones". In: *Optimization methods and software* 11.1-4, pp. 625–653.

Suárez-Fariñas, Mayte et al. (2005). "Harshlight: a corrective make-up program for microarray chips". In: *BMC bioinformatics* 6.1, p. 294.

Subramanian, Aravind et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43, pp. 15545–15550.

Supek, Fran et al. (2011). "REVIGO summarizes and visualizes long lists of gene ontology terms". In: *PloS one* 6.7, e21800.

Takahashi, Kazutoshi and Shinya Yamanaka (2006). "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors". In: *cell* 126.4, pp. 663–676.

Tan, Mingkui et al. (2014). "Riemannian Pursuit for Big Matrix Recovery". In: *ICML*. Vol. 32, pp. 1539–1547.

Ting, Kai Ming (2011). "Sensitivity and specificity". In: *Encyclopedia of Machine Learning*. Springer, pp. 901–902.

Toh, Kim-Chuan, Michael J Todd, and Reha H Tütüncü (1999). "SDPT3—a MATLAB software package for semidefinite programming, version 1.3". In: *Optimization methods and software* 11.1-4, pp. 545–581.

Torres-Sosa, Christian, Sui Huang, and Maximino Aldana (2012). "Criticality is an emergent property of genetic networks that exhibit evolvability". In: *PLoS computational biology* 8.9, e1002669.

Townsend, James, Niklas Koep, and Sebastian Weichwald (2016). "Pymanopt: A Python Toolbox for Optimization on Manifolds using Automatic Differentiation". In: *Journal of Machine Learning Research* 17.137, pp. 1–5.

Tropp, Joel A and Anna C Gilbert (2007). "Signal recovery from random measurements via orthogonal matching pursuit". In: *IEEE Transactions on information theory* 53.12, pp. 4655–4666.

Valverde, Sergi et al. (2015). "Structural determinants of criticality in biological networks". In: *Frontiers in physiology* 6.

Vandereycken, Bart (2013). "Low-rank matrix completion by Riemannian optimization". In: *SIAM Journal on Optimization* 23.2, pp. 1214–1236.

Vizcaíno, Juan Antonio et al. (2016). "2016 update of the PRIDE database and its related tools". In: *Nucleic acids research* 44.D1, pp. D447–D456.

Wagner, Avishai and Or Zuk (2015). "Low-rank matrix recovery from row-and-column affine measurements". In: *arXiv preprint arXiv:1505.06292*.

Wang, Rui-Sheng and Réka Albert (2013). "Effects of community structure on the dynamics of random threshold networks". In: *Physical Review E* 87.1, p. 012810.

Waters, Andrew E, Aswin C Sankaranarayanan, and Richard Baraniuk (2011). "SpaRCS: Recovering low-rank and sparse matrices from compressive measurements". In: *Advances in neural information processing systems*, pp. 1089–1097.

Wei, Ke et al. (2016). "Guarantees of Riemannian optimization for low rank matrix recovery". In: *SIAM Journal on Matrix Analysis and Applications* 37.3, pp. 1198–1222.

Wit, Ernst, Edwin van den Heuvel, and Jan-Willem Romeijn (2012). "'All models are wrong...': an introduction to model uncertainty". In: *Statistica Neerlandica* 66.3, pp. 217–236.

Wold, Svante, Kim Esbensen, and Paul Geladi (1987). "Principal component analysis". In: *Chemometrics and intelligent laboratory systems* 2.1-3, pp. 37–52.

Wu, Zhijin and Rafael A Irizarry (2004). "Stochastic models inspired by hybridization theory for short oligonucleotide arrays". In: *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*. ACM, pp. 98–106.

Xia, Jianguo et al. (2012). "MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis". In: *Nucleic acids research* 40.W1, W127–W133.

Yang, Jing et al. (2016). "Inferring the perturbation time from biological time course data". In: *Bioinformatics* 32.19, pp. 2956–2964.

Yu, Min et al. (2013). "Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition". In: *Science* 339.6119, pp. 580–584.

Zhong, Kai, Prateek Jain, and Inderjit S Dhillon (2015). "Efficient matrix sensing using rank-1 gaussian measurements". In: *International Conference on Algorithmic Learning Theory*. Springer, pp. 3–18.

Zhu, Yuelin et al. (2008). "GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus". In: *Bioinformatics* 24.23, pp. 2798–2800.

Ziemelis, Karl et al. (2001). "Complex systems". In: *Nature* 410.6825, pp. 241–241.