Ph.D. thesis

Understanding responses to external stimuli using network-based approaches

submitted by

Frederik Gwinner

on 15.05.2014

Supervisors

Prof. Heribert Hirt,
Dr. Benno Schwikowski

Jury

Dr. Kathleen Marchal (Reviewer)
Dr. Jost Enninga (Reviewer)
Dr. Ghislain Bidaut
Prof. Denis Thieffry
Declaration of Authorship

I hereby certify that the work presented here is, to the best of my knowledge and belief, original and the result of my own investigations, except as acknowledged, and has not been submitted, either in part or whole, for a degree at this or any other University.

Frederik Gwinner

Paris, 15.05.2014
Summary

In the course of my Ph.D. work, I have developed and applied methods making use of network information to advance the analysis of medium- and high-throughput biological data. My thesis comprises three projects.

Identification of additional proteins in differential proteomics using protein interaction networks

In this study, we developed a novel computational approach based on protein-protein interaction networks to identify a list of proteins that might have remained undetected in differential proteomic profiling experiments. Our approach consists of four steps:
(i) Compilation of a human protein-protein interaction network from public databases;
(ii) calculation of interaction scores based on functional similarity of proteins;
(iii) determination of a set of candidate proteins that are needed to efficiently and confidently connect experimentally identified proteins; and
(iv) ranking of the predicted candidate proteins.
When applied to a 2D-DIGE experiment carried out on two sets of samples from human smooth muscle cells, two of the three highest-ranked proteins, beta-arrestin 1, and beta-arrestin 2, were experimentally tested, revealing that their abundance levels were indeed different between the two sets of samples.

Transcriptional regulatory networks underlying plant responses to environmental stresses

Based on publicly available data, measuring the response of Arabidopsis thaliana to a set of abiotic stresses in a time-resolved manner, we applied two complimentary approaches to derive gene regulatory networks underlying the plant’s response to the perceived stresses. The first method published by [Shiraishi et al., 2010] tries to simultaneously cluster genes with similar profiles and derive regulatory rules between identified clusters in the setting of a state-space model. The second method, which we developed based on a publication by [Chechik and Koller, 2009] identifies genes with impulse-like profiles and predicts regulator-target pairs that can be fit with the same impulse response assuming a given time-lag between regulator and target. Two subnetworks for heat and drought stress response based on the impulse model evaluations are currently evaluated by stimulating plants lacking a set of predicted regulators and measuring transcript abundances of their predicted target genes.

Analysis of transcriptional host immune response signatures specific for distinct stages of infection by Shigella flexneri

Bacterial pathogens enter host cells through tightly regulated conserved molecular mechanisms. During the invasion process they localize to different subcellular niches such as
the cell membrane, pathogen-containing vacuoles and the cytosol. The individual subcellular localization changes of the bacteria are instantly sensed by the host cellular immune system, which activates an appropriate and finely tuned immune response. However, how the differential orchestration of the host immune response is determined by the dynamics of bacterial infection, specifically intracellular localization within cellular compartments, is poorly understood. Our collaborators in Jost Enninga’s group at the Pasteur Institute set out to perform a comprehensive study that integrates spatio-temporal information and transcriptomic profiling using the qPCR-based Biomark platform on collected small cell populations and single cells.

We developed an analysis pipeline tailored to the data mentioned above which performs data normalization, quality control, statistical evaluation of differential gene expression and includes a number of visualization capabilities. The results produced by our pipeline indicate that distinct bacterial subcellular localizations significantly correlate with differentially regulated transcriptional signatures of the host inflammatory and cellular stress response. In addition, using bacterial mutants known to impact specific transcriptional host responses we verified our results and gained novel insights into how different bacterial effectors are involved in activation and subversion of various host immune response pathways during pathogen invasion. For this analysis, we developed a set of statistical tests and network-based analysis approaches specifically adapted to single-cell analysis, which enabled us to reveal cell-specific coordinated gene expression.
Acknowledgements

First and foremost, I would like to thank Heribert Hirt and Benno Schwikowski, who have supervised my work, for supporting me and for articulating most welcome and always constructive criticism. I am very grateful for the opportunity to work with and learn from you.

This work would not have been possible without the great people I had the pleasure to collaborate with. I would like to thank Florence Pinet and Adelina Acosta-Martin from the Institut Pasteur in Lille for helping us in the development, interpretation and validation of our computational Steiner tree approach. I thank everyone from Heribert Hirt’s group at the URGV for letting me join your lab meetings and introducing me to the fascinating world of plant biology. I am deeply grateful to Pierre Jacob for having agreed to spend his valuable time on the experimental evaluation of my predictions. Finally, I would like to thank Jost Enninga and his group, especially Juliane Lippmann, for allowing me to work on the analysis of the extraordinary data you have collected and for the patience with which you explained the underlying biological principles to me.

Furthermore, I would like to thank the members of my thesis advisory committee, Kathleen Marchal and Sébastien Aubourg, for their insightful advice, as well as the members of my thesis jury, Kathleen Marchal, Jost Enninga, Ghislain Bidaut and Denis Thieffry for sacrificing their valuable time to evaluate my work.

The members of the Systems Biology Group at the Institut Pasteur have greatly helped me by discussing ideas, giving various pieces of advice and establishing the comfortable, yet productive, atmosphere inhabiting the group.

I would also like to thank the many other Ph.D. students and post-doctoral researchers working at the Pasteur Institute. It was a pleasure to get to know you during my time here. You have broadened both my scientific and personal horizons and I am proud to be able to call so many brilliant people my friends now.

Thank you to the 'lunch gang' for always being on time and never staying for more than just one beer in the few evenings we actually found the time to go out. Although it saddens me that none of you left Paris and it will thus be easy to see you in the future, I know at least that there wont be anyone waiting for me in other parts of the world. That’s reverse logic!

Last, but not least, I would like to thank my parents, my two wonderful sisters and all the rest of my family for supporting me, for being understanding when I failed to call you regularly and for always being there for me when I needed you, not only in these last years but during all my life.
## Contents

1 Introduction  
1.1 Definition of Systems Biology .......................... 1  
1.2 Large-scale data types/measurement technologies ....... 3  
1.2.1 Transcriptomics ........................................ 3  
1.2.2 Proteomics ............................................. 5  
1.2.3 Protein-Protein Interactions ............................ 8  
1.3 Networks in Systems Biology .............................. 10  
1.3.1 Importance of network context ........................ 10  
1.3.2 Methods guided by network information ............... 10  
1.3.3 Network inference methods ............................... 13  

2 Identification of additional proteins in differential proteomics  
2.1 Background ................................................. 19  
2.2 Proteomics paper ........................................... 19  
2.2.1 Abstract .................................................. 19  
2.2.2 Introduction .............................................. 20  
2.2.3 Materials and Methods .................................. 22  
2.2.4 Results .................................................... 27  
2.2.5 Discussion ............................................... 33  

3 Plant responses to environmental stresses .................. 37  
3.1 Background ................................................ 37  
3.1.1 Sensing of and adaptation to environmental stresses .. 37  
3.1.2 Effect of stress combinations ........................... 43  
3.1.3 Available genome-scale data ............................ 43  
3.2 Modeling responses to abiotic stresses .................... 46  
3.2.1 Data pre-processing ..................................... 46
4 Single-cell transcriptomic signatures elicited by *Shigella* infection

4.1 Background

4.1.1 Host-pathogen interactions

4.1.2 Experimental Setup

4.2 Data analysis pipeline

4.2.1 Pipeline components

4.2.2 Single cell data analysis

4.3 Bulk cell results

4.3.1 Host immune responses to distinct stages of *Shigella* infection are distinguishable

4.3.2 *Shigella* induces a complex T3SS-dependent, but effector-independent bystander response

4.3.3 Five main expression profiles shape the immune response to *Shigella* infection

4.3.4 The effector OspF has the strongest impact on immune responses

4.4 Single cell results

4.4.1 Bimodal expression patterns

4.4.2 Single cell correlation analysis

4.4.3 Effect of infection on cell cycle phase distributions

4.5 Synopsis

4.5.1 Summary

4.5.2 Discussion

Bibliography

A Appendix

A.1 Identification of additional proteins in differential proteomics

A.2 Plant responses to environmental stresses

A.3 Single-cell transcriptomic signatures elicited by *Shigella* infection
List of Figures

1.1 The four M’s of Systems Biology ............................... 2
1.2 General workflow for proteomics experiments .................. 6
1.3 Network context as a crucial means to understand systems .......... 10
1.4 Different mechanisms can explain co-expression .................. 13
2.1 Workflow ..................................................... 23
2.2 Steiner-tree approach ......................................... 24
2.3 Steiner tree computed on the SMC data .......................... 29
2.4 Protein-protein interactions and experimental profile of beta-arrestins. .. 31
3.1 Effect of heat and drought stress on US agriculture ................. 44
3.2 Experimental setup used in the AtGenExpress project .......... 45
3.3 Replicate time-course correlation distributions .................. 48
3.4 Venn diagram of gene selections for cold, osmotic and salt stresses .. 50
3.5 Overview of the state-space model defined by Shiraishi et al., 2010 .. 52
3.6 Inferred network for root response to permanent stresses ........ 55
3.7 Fitted regulation matrices for root response to permanent stresses .. 56
3.8 Expression of four clusters taking part in feedback loops ........ 57
3.9 Expression of four clusters differentially influenced by stresses .... 58
3.10 Fitted regulation matrices for shoot response to permanent stresses .. 61
3.11 Venn diagram of gene selections for heat and drought stresses .. 63
3.12 The impulse model and its parameters .......................... 66
3.13 Impulse model fitting robustness ................................ 67
3.14 Concurrent versus separate impulse model fitting ................. 67
3.15 Impulse network for drought in root tissue ..................... 71
3.16 Impulse sub-network for heat in root tissue .................... 72
4.1 Infection of a host cell by Shigella flexneri ........................ 79
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2</td>
<td>Distinct stages of <em>Shigella</em> infection</td>
<td>80</td>
</tr>
<tr>
<td>4.3</td>
<td>Experimental workflow</td>
<td>82</td>
</tr>
<tr>
<td>4.4</td>
<td>Overview of the developed analysis pipeline</td>
<td>83</td>
</tr>
<tr>
<td>4.5</td>
<td>Bulk and single cell data visualization example</td>
<td>86</td>
</tr>
<tr>
<td>4.6</td>
<td>Evaluation of possible expression thresholds</td>
<td>88</td>
</tr>
<tr>
<td>4.7</td>
<td>Cell cycle phase classification</td>
<td>92</td>
</tr>
<tr>
<td>4.8</td>
<td>Identification of localization-dependent transcriptional signatures</td>
<td>94</td>
</tr>
<tr>
<td>4.9</td>
<td>The bystander response signature</td>
<td>95</td>
</tr>
<tr>
<td>4.10</td>
<td>Genes classified into distinct expression profiles</td>
<td>97</td>
</tr>
<tr>
<td>4.11</td>
<td>Effector-dependency of responses to intracellular infection</td>
<td>99</td>
</tr>
<tr>
<td>4.12</td>
<td>Changes in degree of within-pathway correlation during infection</td>
<td>102</td>
</tr>
<tr>
<td>4.13</td>
<td>Correlation networks during WT <em>Shigella</em> infection</td>
<td>104</td>
</tr>
<tr>
<td>4.14</td>
<td>Correlation networks during Δ<em>ospF</em> <em>Shigella</em> infection</td>
<td>105</td>
</tr>
<tr>
<td>4.15</td>
<td>Cell cycle phase changes during infection</td>
<td>107</td>
</tr>
<tr>
<td>A.1</td>
<td>Significance analysis of the computed Steiner tree</td>
<td>141</td>
</tr>
<tr>
<td>A.2</td>
<td>Distribution of biochemical properties for proteins identified by 2D-DIGE analysis and beta-arrestins</td>
<td>142</td>
</tr>
<tr>
<td>A.3</td>
<td>Evaluation of number of root clusters</td>
<td>143</td>
</tr>
<tr>
<td>A.4</td>
<td>Evaluation of number of shoot clusters</td>
<td>144</td>
</tr>
<tr>
<td>A.5</td>
<td>Inferred network for shoot response to permanent stresses</td>
<td>144</td>
</tr>
<tr>
<td>A.6</td>
<td>Separation between sample groups in PCA</td>
<td>150</td>
</tr>
<tr>
<td>A.7</td>
<td>Genes detected as bimodally expressed</td>
<td>151</td>
</tr>
<tr>
<td>A.8</td>
<td>Cell cycle phases predicted on randomly permuted data</td>
<td>152</td>
</tr>
</tbody>
</table>
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Proteins included in the Steiner tree solution</td>
<td>28</td>
</tr>
<tr>
<td>3.1</td>
<td>Correlations between replicate samples</td>
<td>47</td>
</tr>
<tr>
<td>3.2</td>
<td>Correlations between replicate time-courses</td>
<td>47</td>
</tr>
<tr>
<td>3.3</td>
<td>Response onset times in cold, osmotic and salt stresses</td>
<td>51</td>
</tr>
<tr>
<td>3.4</td>
<td>Response onset times in heat and drought stresses</td>
<td>64</td>
</tr>
<tr>
<td>A.1</td>
<td>Genes measured on the BioMark arrays</td>
<td>150</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

This chapter supplies a very brief introduction to the field of Systems Biology. We will give a definition of what Systems Biology stands for in our minds and discuss commonly used experimental data types and their corresponding measurement technologies. Finally, we will show that interpreting systems with the help of networks is a crucial concept of Systems Biology and discuss how different computational methods make use of this concept. Chapters 2 to 4 of this document then discuss the three main projects I have worked on during my Ph.D. studies.

1.1 Definition of Systems Biology

The term 'Systems Biology' was first coined in the 1960s, when theoretical biologists began creating computer-run mathematical models of biological systems [Schneider, 2013]. The development of measurement technologies allowing the creation of genome-scale data sets in the 1990s has highlighted the need for analysis tools tailored to these data sets and has led to a major resurgence of the field. Coherent with the origin of the term, one commonly used definition for Systems Biology is "mathematical modeling of biological systems". This definition does, however, fail to recognize the importance of the experimental side of Systems Biology. In the beginning of the 21st century, Leroy Hood, back then the president of the Institute for Systems Biology in Seattle, identified the integration of data sets measuring different aspects of the cellular state, like DNA, RNA, proteins and metabolites, as paramount to the field of Systems Biology. He furthermore stressed that measurement technology — and the ability to get quantitative data from many different conditions — is the biggest bottleneck in Systems Biology [Henry, 2003]. Douglas Lauffenburger, back then a professor of biological engineering at the Massachusetts Institute of Technology, pointed out the importance of hypothesis generation in Systems Biology. He stated that, faced with highly complex data, modeling of the underlying system has to be predated by figuring out the right questions to ask, i.e. by the generation of testable hypotheses through data mining approaches [Henry, 2003].

It can be argued that, to date, no clearer definition of Systems Biology than the ones made a decade ago has been agreed upon. One of the few points universally acknowledged is,
1.1. Definition of Systems Biology

however, that the goal of Systems Biology is to make better use of genome-scale data sets and push analyses beyond the purely descriptive stage towards predicting the behavior of biological systems. These genome-scale, commonly referred to as "omics", data sets measured by high-throughput experimental technologies posed, and still do pose, considerable challenges due to a number of factors we will elaborate on in the next section. To me personally, Systems Biology is thus characterized by a tight link between computational approaches and the experimental measurements and can be defined by its two main tasks: (i) to aid the observation of biological systems by extracting the highest possible degree of information from complex large-scale data and (ii) to develop mathematical models of biological systems targeted towards the generation of testable hypotheses. Chapters 2 and 4 represent methods developed to aid in the observation and analysis of biological systems, while Chapter 3 describes two modeling approaches.

Consequently, computational Systems Biology approaches can never be regarded separately from experimental biology but should be embedded in a cycle containing experimentation and computational modeling components. Douglas Lauffenburger refers to this as the "four M's" of Systems Biology: measurement, mining, modeling, and manipulation (c.f. Figure 1.1). These four M's are part of an iterative process, beginning with manipulating the system. Once a system is perturbed, it is measured using a high-throughput, multivariate technology. The data are then mined to elucidate hypotheses that, when cast in terms of formal computational models, form the basis for a new manipulation of the system [Henry, 2003].

Figure 1.1: Systems Biology as a cycle involving the 4 M's: measurement, mining, modeling, manipulation. This figure was taken from Figure 1 of [Ideker et al., 2006].
1.2 Large-scale data types/measurement technologies

In the following, we will describe large-scale data types analyzed in the course of my PhD work and discuss the strengths and biases of different corresponding measurement technologies.

1.2.1 Transcriptomics

The term transcriptomics refers to the measurement of RNA concentrations present in biological samples. While many different types of RNA can be measured using transcriptomic technologies, we focus here on the measurement of messenger RNA (mRNA) concentrations which are often used as a proxy for protein concentrations in modeling.

Microarray technologies

Techniques to measure RNA concentrations have been around for a long time. The basic principle behind those techniques is the immobilization of DNA sequences (called probes) specific for the transcripts of interest on either membranes (as in northern blots) or solid surfaces. The extracted RNA samples are then reverse transcribed into DNA and labels (earlier radioactive labels, by now fluorescent labels) are added. The labeled DNA is allowed to hybridize to the immobilized probes and, after a washing step, only DNA bound to the probes is left and the label intensity can serve as a read-out for transcript concentration. Initially, long complementary DNAs (cDNA)s immobilized on glass slides were used as probes (e.g. [Schena et al., 1995]). Later, new oligonucleotide array designs were developed, which contained shorter (~20 bases) DNA probes, that were synthesized directly on a solid surface using photolithographic technologies. The photolithographic synthesis allowed miniaturization of the arrays and consequently the spotting of considerably more probes per array. Individual probes matching different sequence stretches of a transcript of interest are typically scattered over the array to cancel local differential hybridization efficiency effects. Read-outs from the individual probes matching a given transcript (so-called probe sets) can then be combined computationally into a single fluorescence signal that correlates the abundance of a given transcript. Oligonucleotide arrays first allowed to measure all transcripts of simple eucaryotic organisms [DeRisi et al., 1997] and significant portions of the human transcriptome [Lockhart et al., 1996].

While they represent a great improvement over cDNA-based arrays, it has been shown that oligonucleotide array data still presents a very noisy estimate of actual RNA concentrations, which hinders especially the accurate estimation of low RNA concentrations. Causes of this noise are technical factors related to the microarray, like e.g. variations in the amount of DNA spotted or the probe sequence design, as well as technical noise produced by the image analysis procedure, which involves e.g. image segmentation. Other factors that influence the measured fluorescence intensities are biological or chemical in nature. Examples are variations in the experimental setup like temperature or the concentration of buffers used to wash the arrays [Barash et al., 2004]. Due to these factors, direct comparison of individual transcripts measured on a chip and — to an even greater degree
1.2. Large-scale data types/measurement technologies

— comparison of transcript levels detected on different arrays is not reliably possible on the directly measured fluorescence intensities.

To make arrays more comparable, a number of microarray normalization methods, like e.g. the Affymetrix Microarray Suite (MAS, [Hubbell et al., 2002]), the Robust Microarray Analysis (RMA [Irizarry et al., 2003]) and a variant of RMA correcting for estimated differences in probe binding affinities due to GC content (GCRMA [Wu and Irizarry, 2007]), have been developed. It has been shown that the employed normalization method can have profound effects on biological findings like e.g. the detection of differentially expressed genes [Hoffmann et al., 2002]. Although a number of studies have been carried out to evaluate the performance of different normalization methods (e.g. [Barash et al., 2004]; [Harr and Schloetterer, 2006]; [Lim et al., 2007]), they have drawn partially inconsistent conclusions and no universally accepted best-practice normalization method has been agreed upon yet.

Even the newest array platforms require a rather large amount of total RNA input (100 ng according to an Affymetrix protocol for sample preparation [http://media.affymetrix.com/support/downloads/manuals/3_ivt_express_kit_manual.pdf]). They thus do not allow the study of small quantity samples as e.g. mRNA samples derived from single cells.

PCR-based technologies

The polymerase chain reaction (PCR) is a method to amplify specific DNA sequences initially developed in the 1980s by Kary Mullis [Bartlett and Stirling, 2003]. It consists of a repeated application of two steps: The first step consists of melting double-stranded DNA (dsDNA) into its two complementary single-stranded DNA (ssDNA) molecules by subjecting the sample to high temperatures. In the second step, the sample is brought back to a lower temperature and a thermotolerant DNA polymerase complements and thus amplifies selected ssDNA molecules. The selection of amplified DNA is achieved by adding short oligonucleotides, so-called primers, that are complementary to the 3’ end of the desired DNA and facilitate polymerase-based complementation of the ssDNA into dsDNA. The two steps form one cycle of the PCR and are repeatedly applied in order to specifically amplify a desired DNA molecule.

Quantitative or real-time PCR (qPCR) was designed to measure the quantity of a given RNA molecule in a sample. To this end, RNA molecules are first reverse-transcribed into cDNA and then subjected to a PCR. After each cycle of this PCR, qPCR additionally measures the concentration of the amplified DNA molecule, by using fluorescence markers only active when integrated in dsDNA (typically either intercalating agents or fluorescence-labeled nucleotides). The measured fluorescence intensity increases after each cycle and the cycle at which a baseline threshold of intensity is exceeded serves as a measurement of the initial cDNA concentration in the sample. As each cycle leads to a duplication of the initial amount of cDNA, the measured cycle threshold (CT) values are inversely proportional to the the logarithm to the base 2 of the initial RNA/cDNA concentration.
1.2. Large-scale data types/measurement technologies

Historically, qPCR was mostly applied on a small number of transcripts, as each transcript quantification has to be run in a separate PCR. The technological improvement in the manufacturing of microfluidic devices recently allowed the extension of qPCR to medium-throughput scale. To this end, arrays consisting of multiple small PCR reaction chambers into which multiple samples and multiple primer sets can be loaded selectively, have been developed. BioMark arrays designed by the company Fluidigm are one example of such microfluidic devices. They consist of 96x96 PCR reaction chambers, each of which can be loaded with a specific sample/primer combination, thus allowing the simultaneous measurement of 96 genes in 96 samples. Specifically designed array readers and quantitation software then interprets fluorescence read-outs and estimates one CT value per chamber.

qPCR offers much more accurate and sensitive detection of RNA concentrations than microarray technologies and needs as little as 10 ng of total RNA as input material [Jang et al., 2011]. Thus, it allows the measurement of transcript abundances in samples derived from single cells (e.g. [Citri et al., 2012]). It is, however, also subject to noise and variations in the detection of different transcripts. Key factors include primer efficiency, non-specific amplification, as well as potential mechanical problems during the loading of the chambers. The exponential nature of the amplification procedure furthermore leads to less precise estimation of lowly concentrated RNAs in comparison to more abundant ones.

1.2.2 Proteomics

The term proteomics describes the holistic measurement of proteins present within, e.g., a cell or a tissue. Proteins encode the actual molecular machinery of living cells and are much more diverse in terms of their physico-chemically properties than DNA or RNA molecules [May et al., 2011]. Additionally, they can be subjected to a number of modifications like e.g. phosphorylation, methylation and ubiquitination, which can influence their functional activity by inducing changes in subcellular localization, stability, binding affinity to DNA or other proteins and enzymatic activity. Due to the physico-chemical diversity of proteins, the proteome, defined as a snap-shot of proteins expressed in a cell or tissue at a defined time point [Wilkins et al., 1996], is much harder to study than genome or transcriptome.

Technologies developed to study the proteome encompass methods to qualitatively or quantitatively measure protein abundances as well as measure differences in abundance between two samples. They typically consist of two steps: in the first, separation step, the proteins contained in a complex sample are ordered according to properties like e.g. protein size or their isoelectric point, the pH at which the protein carries no charge. In the second step, individual proteins or sets of proteins with similar separation properties are broken down into smaller peptides through digestion and then subjected to identification typically using mass spectrometry (MS). Proteomics technologies can be broadly categorized in two classes: gel-based and gel-free methods (c.f. Figure 1.2).
1.2. Large-scale data types/measurement technologies

**Gel-based methods** separate proteins by running them through a porous gel in an electrophoresis step, which allows separation by protein size, since large proteins travel slower than small ones. In 2D-gels, the proteins are first separated according to their isoelectric point (for example on a pH gradient) and then loaded onto a gel and separated in the second dimension according to protein size. Bands or spots containing proteins in the gel are identified and the quantity of protein in each spot estimated (often with the help of image analysis tools). To identify proteins contained in each spot, they are excised and subjected to identification in a mass spectrometer. For differential proteomics the two-dimensional difference in-gel electrophoresis (2D-DIGE, [Unlü et al., 1997]) method, which consists of running two samples labeled with different fluorescent dyes in a single 2D gel, has allowed more accurate sample comparisons by avoiding the effect of gel-specific variations. MS-based methods have initially been developed for protein identification and not quantification. Consequently, they are commonly applied for protein identification after estimation of protein quantity using a gel-based method.

![Figure 1.2: General workflow for proteomics experiments: This figure was taken from Figure 1 of May et al., 2011](image-url)
1.2. Large-scale data types/measurement technologies

**LC-MS/MS** In a typical MS experiment, proteins digested into peptides are subjected to liquid chromatography (LC) followed by identification or quantification by MS. The basic principle of LC is to separate solute analytes (peptides) in a fluid that flows over a column containing solid particles to which the individual analytes adhere with different strengths and are thus retained for longer or shorter periods of time in the LC column. The outflow of the LC column is then injected into a mass spectrometer.

The mass spectrometer ionizes the analytes and then separates them further according to their mass over charge ratio (m/z). To this end, various technologies have been developed for the separation of ions, like e.g. by their time of flight (TOF) or by their deviations in magnetic fields. The relative abundance of ions with a given m/z is recorded in a spectrum and these spectra can be matched against spectral databases to identify individual peptides and proteins present in the analyzed sample. Newer generations of mass spectrometers are additionally capable of sorting out ions of a given m/z and can be used in so-called MS/MS experiments, where precursor ions with a specific m/z are enriched, further fragmented, and subjected to a second round of MS allowing to predict the aminoacid sequence of precursor peptides from the recorded m/z values of their fragments.

**Gel-free quantitative methods** are typically based on the LC-MS/MS technology described above. Absolute quantification of peptides by LC-MS/MS is hampered by the fact that physico-chemical properties of peptides influence the intensity of the response in a mass spectrometer. Thus, absolute quantification of a single sample requires the use of standards, which consist of peptides labeled with stable isotopes. The mass spectrometer is then able to recognize the difference in mass between a peptide and its labeled counterpart in the standard and the absolute quantity of the peptide can be inferred by comparing its measured signal intensity to the one of the standard. For the case of differential proteomics, a similar approach can be taken in which the two samples to be compared are subjected to differential stable isotope labeling to create specific mass tags that can be recognized by the mass spectrometer and at the same time provide the basis for quantification [Bantscheff et al., 2007]. Additionally, so-called label-free methods attempt to directly compare two or more LC-MS/MS experiments by computationally matching peaks detected in the experiments to each other and comparing the respective measured signal intensities. Due to technical differences between individual experiments and limits in terms of the m/z detection accuracy of mass spectrometers, the matching procedure is far from trivial and typically requires using additional information, like measured retention times in the LC column ([Strittmatter et al., 2003]; [Vandenbogaert et al., 2008]).

Although proteomic technologies have greatly evolved during the past decade, they still confront challenges, mainly in terms of sensitivity of detection. 2D-DIGE is, additionally, limited in the detection of proteins with extreme molecular weight, extreme pI or high hydrophobicity due to particularities of the gel-based separation step [Friedman et al., 2009]. Gel-free methods like LC-MS/MS have initially been developed for protein identification and not quantification and are subject to technical variations [Prakash et al.,]
1.2. Large-scale data types/measurement technologies

and data analysis inconsistencies [Bell et al., 2009]. Existing methods to render MS-based approaches quantitative introduce further sources of experimental variation leading to quantification errors [Bantscheff et al., 2007] and also show limitations in the detection of lowly abundant proteins, especially in complex samples.

1.2.3 Protein-Protein Interactions

Many molecular functions of a cell are exerted not by single proteins but rather by protein complexes. Stable protein complexes but also transient protein-protein binding are necessary for DNA, RNA and protein synthesis, proper protein folding, degradation of proteins, signal transduction, many enzymatic reactions and various other cellular functions. Consequently, large-scale maps of protein-protein interactions — also called interactomes — have been described as crucial parts of the wiring diagrams of cells [Tan et al., 2007]. Furthermore, since interacting proteins tend to play a role in similar biological processes, protein interaction networks have been used in inferring functions of previously uncharacterized proteins [Sharan et al., 2007].

The most prominently used large-scale method to detect protein-protein interactions are yeast two-hybrid (Y2H) screens based on a method first described by [Fields and Song, 1989]. Y2H screens make use of the architecture of the yeast GAL4 protein, which is essential for yeast to grow on galactose medium. GAL4’s function depends on the spatial proximity of two of its domains. The screen works by expressing two proteins each fused to one of the GAL4 domains. If the proteins are able to bind to each other, GAL4 activity is restored and the yeast can grow, otherwise growth is inhibited by the lack of galactose utilization capabilities. As the gene fusion and read-out are relatively easy to automate, Y2H screens have been applied to elucidate large portions of the interactomes of many organisms including yeast, worms, flies, plants and humans [Gingras and Raught, 2012]. Detection of an interaction in Y2H does, however, not necessarily prove interaction of the proteins in the actual organism of interest, where tissue specificity, subcellular localization and post-translational modifications of proteins can additionally influence binding behavior.

Another prominent experimental method for detection of protein interactions is tandem affinity purification (TAP) followed by mass spectrometry [Rigaut et al., 1999]. This approach is targeted towards finding all proteins, directly or as part of a larger complex, bound to a protein of interest. To this end, sequence tags are fused to the bait protein, protein complexes are cross-linked and complexes involving the bait protein are purified using agents specifically binding the sequence tags. Then, the proteins forming a complex with the bait are identified using LC-MS/MS. Due to the experimental setup, TAP-MS cannot identify binary interactions, but has the advantage of measuring complexes actually occurring in the organism of interest. Applying TAP-MS in an automated way has been successfully applied to elucidate a yeast interactome map [Krogan et al., 2006].

As showcased by the two methods described above, different technologies for measuring protein interactions have different strengths and weaknesses and consequently experi-
1.2. Large-scale data types/measurement technologies

Experimental evidence from different methods is overlapping poorly [von Mering et al., 2002]. The integration of distinct experimental evidence for interaction in meta-databases has been shown to improve the quality of interactome maps ([Ramírez et al., 2007]; [Rivas and Fontanillo, 2012]). Nevertheless, to date, protein-protein interaction maps for higher organisms are still static and do not reflect dynamic changes in protein binding as no experimental method capable of measuring all interactions occurring at a given time in a cell is available. Furthermore, protein interaction maps for higher organisms are far from being complete and show a bias towards interactions involving highly expressed and highly studied proteins [Ivanic et al., 2009].
1.3 Networks in Systems Biology

Since Systems Biology can be defined as the study of biological systems with the help of computational means, networks, either as an input used to guide computational approaches or as the output of computational modeling, play a key role in it.

1.3.1 Importance of network context

Figure 1.3 describes a toy example showcasing the importance of including network context when studying biological systems. Imagine a pathway composed of five proteins A to E, with two of them, B and D being subject to a posttranslational modification that positively affects their activity (Figure 1.3A). Assume further that the regulatory network structure displayed in Figure 1.3A underlies the studied pathway. Using reductionist approaches targeted at the study of the individual proteins B and D, understanding the effect of these modifications on the whole pathway is impossible. In fact, detailed studies of protein B would probably predict it to not be involved in the pathway. Only when studying all five proteins of the pathway together and making measurements of their activity changes after perturbations of the system, one would be able to understand that B’s perceived lack of effect on the pathway state is caused by a negative feedback mechanism via C and A.

![Figure 1.3](image)

Figure 1.3: A toy example showing the importance of studying biological systems by looking at network context: please refer to the accompanying text for an explanation.

1.3.2 Methods guided by network information

Protein-protein interaction networks have been shown to contain information about protein functions. This is due to the modular architecture underlying the molecular machinery of living systems [Barabási and Oltvai, 2004], composed of proteins that form
1.3. Networks in Systems Biology

relatively static complexes, like e.g. the ribosome as well as dynamically changing complexes like e.g. in signal transduction pathways consisting of protein kinase cascades.

The 'guilt by association' principle states that proteins sharing common features are likely to have similar functions and is commonly used in computational methods for protein function prediction. Previously, such methods were mainly based on features derived from proteins' biochemical properties, their sequence [Friedberg, 2006] as well as their structure [Domingues and Lengauer, 2007]. By defining similarity measures on such features, annotated proteins similar to a protein of interest can be found and machine learning methods used to decide whether their functional annotations can be transferred (as e.g. in [Weinhold et al., 2008]). The 'guilt by association' principle has, however, also been extended to predict protein function through proximity in protein interaction networks. Two main principles can be distinguished here: direct methods use functional annotations enriched in the network neighborhood around a protein of interest, while module-assisted methods first identify modules of related proteins — typically by applying clustering approaches — and then annotate each module based on the known functions of its members [Sharan et al., 2007].

But large-scale network data has been proven useful not only to the functional annotation of proteins. A large number of computational approaches are guided by network data of different kinds in various ways. Listing all of these approaches would go beyond the scope of this introduction. Instead we will present applications of network-guided computational tools on the basis of a few examples.

Sample classification [Hofree et al., 2013] published a method for cancer sample stratification, called network-based stratification (NBS). They show that standard clustering approaches applied to somatic mutation profiles obtained by exome sequencing are unable to distinguish between cancer sub-types with different causes and clinical outcomes. This is due to the low overlap in individual mutations between individual cancer samples. Spreading of mutation evidence over interaction networks is able to abstract from individual single nucleotide polymorphisms to network regions affected by mutations and allows a better comparison between samples. The authors show that such network smoothing (using a network propagation algorithm based on random walks) in conjunction with a clustering method (non-negative matrix factorization) is able to classify human cancer samples into sub-classes which show different clinical outcomes and correspond well to histological classifications. Furthermore, they derive predictive signatures for the individual sub-types by looking at differentially affected network regions. Their network smoothing approach was applied to protein-protein interaction networks (STRING), interaction networks based on integrating various experimental evidence (HumanNet) as well as networks focused on canonical pathways (PathwayCommons). Interestingly, after filtering of low-confidence interactions, all three tested networks led to similar results.

A supervised learning method for prediction of sample classes, called network-guided forests has been published by [Dutkowski and Ideker, 2011]. It is based on training a sample class predictor on data with known classification. The training step is guided by
1.3. Networks in Systems Biology

...a physical interaction network between transcription factors. It constructs gene modules (i.e., connected components) in the network by iteratively constructing a decision tree, in each step restricting the feature selection to neighbors of previously added genes. Each of the learned decision trees can be seen as a set of logical rules determining the activity of its corresponding gene module. The outputs of the individual decision trees are then integrated to predict a sample’s class. The method does not only allow sample class prediction for unclassified samples, but also identifies modules and the regulatory rules governing their activity. This allows postulating hypotheses about the involvement of the predicted modules in causal mechanisms behind phenotypic differences. The authors show that their method predicts robust modules that capture causal mechanisms underlying developmental stages and cancer progression.

Candidate prioritization [Verbeke et al., 2013] published EPSILON, a framework for candidate gene prioritization in eQTL studies. Expression quantitative trait locus (eQTL) studies identify loci on the genome whose genetic variation within a population is associated with expression variability of target genes. The main application of eQTL studies is the identification of novel marker genes, whose mutations are causal for diseases. Since the identified loci are typically derived from relatively few genetic markers, they encompass multiple candidate genes that could be causal for the observed target gene variability. Thus, candidate prioritization is needed to minimize the number of genes to be tested in follow-up experiments. A number of approaches using protein interaction networks have been developed for this task. The common idea is to derive a network-based similarity measure between candidate genes and the target gene and to prioritize candidates with high similarities. The presence of hub nodes in global interaction networks, which exhibit a high number of connections to other nodes, can heavily influence such similarity measures. This is especially detrimental in the case of so-called "promiscuous proteins", which bind many functionally distinct proteins and thus offer little information about true biological similarity of genes. EPSILON thus first constructs a local network around the candidates and the target. This is achieved by selecting a set of shortest paths between each candidate and the target gene. The authors show that a number of previously published network-based similarity measures achieve better prioritization on such local networks as compared to applying them on the global interaction network. EPSILON is implemented as a modular framework and thus allows the combined use of different eQTL mapping methods and network similarity measures.

Proteomics data interpretation A number of methods using interaction networks in the downstream analysis of proteomics data sets have been developed as well. In chapter 2.2, we describe our recently published computational approach to predict proteins that have potentially escaped identification in gel-based differential proteomics experiments and discuss a few other methods using similar approaches to facilitate proteomics data interpretation.
1.3. Networks in Systems Biology

1.3.3 Network inference methods

Since the onset of high-throughput measurement technologies, a myriad of methods for the prediction of networks underlying biological systems have been developed. They can be distinguished according to the type of predicted network (e.g., regulatory interactions, protein-protein, protein-DNA/RNA interactions, metabolic reactions, signaling events), according to the type of data the predictions are based on (e.g., static gene expression, time-resolved gene expression, (phospho-)proteomics data, transcription factor binding, text-mining of scientific literature and combinations thereof) and according to the mathematical principles used. Congruent with the scope of this thesis, we will focus on methods inferring regulatory networks based on transcriptomic data, discuss mathematical principles commonly used in this flavor of network inference and discuss their respective strengths and limitations.

Inference on static expression data

Co-expression networks The first attempts at discovering cellular networks from measured expression data have focused on the analysis of expression patterns over multiple samples. If two genes show similar changes in expression, as measured e.g. by correlation coefficients [Stuart et al., 2003], they can be inferred to likely be functionally related to each other. Such co-expression networks especially when computed over large collections of microarray data measured under different conditions can be very useful for protein function prediction and detection of co-regulated modules. Correlation measures are, however, inherently unable to distinguish between direct and indirect regulatory interactions. Figure 1.4 shows this based on a simple example. An observed co-expression between the genes X, Y and Z can be explained by different regulatory pathways (1, 2) between the genes or can be due to a co-regulation of the three genes by an additional entity H (3). To make assumptions about which interactions are most likely to actually occur, co-expression based methods either use assumptions about typical characteristics of regulatory networks or make use of additional information like e.g. known transcription factor activity of a given gene. Due to their simplicity, co-expression networks can

![Figure 1.4: Different mechanisms can explain co-expression: The plot in the dashed box shows a co-expression graph. Regulatory networks 1-3 show possible explanations for the observed coexpression. This figure was taken from Figure 1 of [Markowetz and Spang, 2007]](image-url)
be inferred easily on genome-scale data, but also are not able to model complex behavior, like e.g. cooperativity between transcription factors.

**Conditional independence** To alleviate the shortcomings of co-expression methods in predicting direct regulatory interactions, methods based on the mathematical notion of conditional independence have been developed. The basic principle behind those methods is to assess the conditional independence of genes A and B given a background C, or in a formula: \( A \perp B \mid C \). Coming back to the example used in the previous paragraph, the regulatory networks 1 and 2 can then be distinguished by asking the question whether, knowing the state of \( Y \), the state of \( Z \) is independent of the state of \( X \). There are different model classes based on this conditional independence formulation which differ in terms of what background \( C \) is considered. In the following, we will list these model classes in order of increasing complexity.

**Low-order dependence models** ask the question whether two genes are conditionally independent given any single other gene. The Algorithm for the Reverse Engineering of Accurate Cellular NEtworks (ARACNE, [Basso et al., 2005]) uses such first order conditional independence by computing conditional mutual information. The Context Like-lihood of Relatedness (CLR, [Faith et al., 2007]) uses a similar approach in which the mutual information (MI) between a potential regulator and its target are compared to a background distribution constructed over all MI values measured in pairs involving either the regulator or the target. This procedure assures that high mutual information due to systemic differences between sample groups is canceled out and is able to prioritize specific regulator target pairs over "promiscuous" regulators and targets.

**Full conditional models**, like e.g. Gaussian graphical models (GGMs) predict regulations between genes that are conditionally dependent given all other measured genes as indicated by non-zero partial correlations. Since full conditional dependence is hard to estimate on data with few samples and many genes, sparseness assumptions or heuristics have to be employed in the estimation of these models [Markowetz and Spang, 2007]. An example of a GGM is the GeneNet model published by [Opgen-Rhein and Strimmer, 2007].

Using the concept of conditional independence as described above helps distinguishing between direct and indirect regulatory interactions. It does, however, still not infer the directionality of regulatory interactions (i.e. which of the interacting genes regulates the other). Most of the above mentioned methods, like in the case of co-expression networks, thus restrict the inference to gene pairs containing at least one gene known to act as a transcription factor.

Finally, **Bayesian networks** can be conceptually seen as looking for gene pairs that are conditionally dependent given all possible combinations of other genes. This is realized by restricting the inferrable networks to the class of directed acyclic graphs (DAGs). In such a DAG, each node has an attached probability distribution conditional on its parents. Bayesian networks are generally inferred in three steps: in the model selection step, a set of DAG structures is proposed. The parameter fitting step tries to find parameters for the individual conditional probability distributions that fit well with the observed data.
1.3. Networks in Systems Biology

and the model with the highest fitness is then selected as the inferred network structure. Of these three steps the model selection is the most critical, as even for few genes, an exhaustive exploration of all possible DAGs becomes unfeasible. The use of DAG models has the inherent advantage to offer a directed representation of the inferred regulatory rules. Prominent examples of Bayesian network inference methods are \cite{Friedman2000} and the tool Banjo \cite{Hartemink2001}.

Due to the increasing complexity of the discussed concepts, the resulting models are able to capture increasing levels of complexity in terms of a systems' behavior. This does, however, come at a prize. While ARACNE and CLR for example have been applied to genome-wide data sets for prokaryotes, GGMs and Bayesian network models are typically restricted to hundreds of genes (800 in \cite{Opgen-Rhein2007} and \cite{Friedman2000}) even for data sets with many samples.

Other models Some network inference models do not readily fit in the classification used above. The module networks method \cite{Segal2003}, for example, searches for sets of genes whose expression variations can be explained by the expression levels of few potential regulators. The regulation of each module is modeled as a tree. The learning procedure starts with a clustering of genes into modules. Then, the learning of module-regulating trees and re-assignment of genes to modules are optimized iteratively using an Expectation-Maximization (EM) procedure. The network inference tool LeMoNe \cite{Bonnet2010} is based on the idea presented in \cite{Segal2003}, but uses ensemble-based probabilistic optimization techniques instead of an EM-algorithm and has been applied successfully to various biological systems. Such module-based models have the advantage of splitting the problem of inferring regulatory rules on the whole system into smaller parts, namely the modules, which allows the inference of relatively complex regulatory rules for genome-scale systems.

Inference on time-course expression data

Modeling on static expression data inherently assumes that the observed biological system is at a steady-state, i.e. a situation in which the expression of all genes is stable and will not change until a perturbation of the system occurs. Data measuring a biological system during changes from such steady-states, e.g. due to external stimuli, can offer additional information about regulatory interactions. To model regulatory networks underlying such kinetic behavior of a system over time, another set of modeling approaches has been developed.

Time-lagged correlation Time-series measurements allow to evaluate the temporal order of deregulation events. This is based on the credo that "cause comes before consequence", i.e. a change in a regulator's expression should predate a similar change in its target gene. In its simplest form, akin to co-expression networks on static data, the measurement of time-lagged correlations between pairs of genes allows making an inference of the order of deregulation events. As in the case of co-expression networks, such methods can easily be applied to large-scale data, but offer relatively little detail. Furthermore,
the number of situations to be evaluated is increased in comparison to static data, as multiple possible time-lags can be considered for each gene-pair.

**Granger causality models**  The concept of Granger causality [Granger, 1980] is based on finding genes $x$, whose expression value at a given time $t$ harbors more information about future expression values of a second gene $y$ than already contained in $y$’s expression up until the time point $t$. The simplest methods based on Granger causality define a Granger causality index between two genes by fitting pairwise autoregressive (AR) models (e.g. in [Mukhopadhyay and Chatterjee, 2007]). More evolved methods, akin to the concept of conditional independence on static data, use measures of conditional or partial Granger causality (as e.g. lined out in [Hempel et al., 2011]).

The notion of Granger causality has been developed in the field of econometrics, where time-series data with many data points are typically available. Consequently, comparative studies have shown that these methods do not translate very well to the short time-series data (on average less than ten measured time points [Ernst and Bar-Joseph, 2006]) typical for biological data sets (Zou et al., 2009; Hempel et al., 2011). This is further corroborated by the fact that the few successful applications of Granger causality methods to biological data have been reported on small systems underlying processes which are clearly governed by cyclic temporal expression patterns like e.g. a cell cycle circuit [Guo et al., 2008] or an already well-known small network of genes involved in diurnal processes in plants [Zou et al., 2009].

**Kinetic models**  Kinetic models directly represent regulatory interactions by a set of equations describing the rates of expression change of a given gene $x_i$ dependent on the expression of other genes $x_j$. The functions used to describe these dependencies can be either linear combinations (linear differential equations) or more complex (non-linear differential equations). As the measured kinetics are discrete, the corresponding models are typically adapted to discrete time models (sometimes called difference equations). A typical linear difference equation model is for example given by:

$$x_i(t + 1) = \sum_{j=1}^{N} A_{i,j} \ast x_j(t) + b_i u \quad (1.1)$$

with a matrix $A$ describing individual influences on $x_i$ exerted by all other genes $x_j$ and the vector $b$ representing the influence of an external perturbation $u$.

The advantage of kinetic models is that, once fit to the data, they allow quantitatively predicting the effect of network perturbations, like e.g. knocking out a given gene, as well as extrapolation of the systems behavior after the last measured time point. Both linear and to an even greater extent the systems behavior after the last measured time point. Both linear and to an even greater extent non-linear difference equations require, however, the fitting of a large number of parameters ($N^2$ parameters in the matrix $A + N$ parameters in the vector $b$ for the above linear difference model). A number of methods have been developed to reduce the number of parameters by e.g. enforcing a sparsity constraint on the regulation matrix. Despite such efforts, kinetics models have mainly been restricted to small systems (e.g. 9 genes of the DNA damage response pathway of *E.Coli* [Bansal et al., 2006] or detailed modeling of the EGF-receptor system [Wiley et al., 2003]).
Data simplification As laid out in the previous paragraphs, methods for network inference on time-resolved data are less well established than the ones developed for static expression data [Bar-Joseph, 2004]. The limited number of time points available in biological data makes the problem very challenging, especially when considering large and complex systems. One solution to this is to "increase" the number of time points through interpolation, but such approaches tend to overfit the individual noisy gene expression data [Bar-Joseph, 2004] or require restricting the shape of interpolated curves according to prior assumptions (as e.g. done by [Chechik and Koller, 2009]). Furthermore, a number of simplification schemes have been applied to time-series data to make the problem of network inference more tractable. Simplification on the level of expression measurements is, for example, achieved by data discretization into binary or nominal values, which then allows qualitative modeling of the system [Wang et al., 2008]. Other approaches simplify the system to be inferred by grouping similar genes into clusters and inferring regulatory interactions between these gene groups (as e.g. done in a state-space model framework by [Hirose et al., 2008]).

As laid out in this section, the inference of regulatory networks from genome-scale expression data is a very challenging task. The Dialogue on Reverse-Engineering Assessment and Methods (DREAM [Stolovitzky et al., 2007]) conferences, have led to the annual creation of new network inference challenges in which participants can apply their individual methods to data supplied by the organizers and central evaluation of the results is carried out. The results have initially been very sobering ("reliable network inference from gene expression data remains an unsolved problem" DREAM3 [Marbach et al., 2010]), but have shown considerable progress in that the DREAM5 competition revealed methods to be able to predict genome-scale networks for prokaryotes with 50% precision (DREAM5 [Marbach et al., 2012]). A general observation made in all iterations of the DREAM challenges is that combining the results of individual methods largely outperformed the best single method.
1.3. Networks in Systems Biology
Chapter 2

Identification of additional proteins in differential proteomics

2.1 Background

This project was carried out in collaboration with a research group at the Pasteur institute in Lille, lead by Dr. Florence Pinet. Her group’s main research focus is on studying cardiovascular diseases, like aortic aneurysms, chronic cardiac insufficiency and myocardial infarction. They develop and utilize differential proteomics approaches to screen for candidate genes explaining susceptibility of individual patients to the aforementioned cardiovascular diseases by analyzing clinical samples obtained from the local hospital. Adelina Acosta-Martin, at the time a post-doctoral student in Dr. Pinet’s group, carried out the experimental work and was closely involved in the development and validation of our computational approach. The following sections contain a reproduction of our article entitled "Identification of additional proteins in differential proteomics using protein interaction networks" published in the journal Proteomics in April 2013. All mentioned experimental studies were conceived and conducted by Florence Pinet and Adelina Acosta-Martin. The computational approach and computational validation studies were conceived and implemented by Benno Schwikowski and me. Consequently, detailed experimental procedures were moved from the main text of the paper to the appendix of this document and only the experimental studies necessary for the understanding of the paper were kept in the main manuscript. The full manuscript is freely available online at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3964363/

2.2 Proteomics paper

2.2.1 Abstract

In this study, we developed a novel computational approach based on protein-protein interaction networks to identify a list of proteins that might have remained undetected
in differential proteomic profiling experiments. We tested our computational approach on two sets of human smooth muscle cell (SMC) protein extracts that were affected differently by DNase I treatment. Differential proteomic analysis by saturation DIGE resulted in the identification of 41 human proteins. The application of our approach to these 41 input proteins consisted of four steps: (i) Compilation of a human protein-protein interaction network from public databases; (ii) calculation of interaction scores based on functional similarity; (iii) determination of a set of candidate proteins that are needed to efficiently and confidently connect the 41 input proteins; and (iv) ranking of the resulting 25 candidate proteins. Two of the three highest-ranked proteins, beta-arrestin 1, and beta-arrestin 2, were experimentally tested, revealing that their abundance levels in human smooth muscle cell samples were indeed affected by DNase I treatment. These proteins had not been detected during the experimental proteomic analysis. Our study suggests that our computational approach may represent a simple, universal, and cost-effective means to identify additional proteins that remain elusive for current 2D gel-based proteomic profiling techniques.

### 2.2.2 Introduction

Proteins and their abundance or state changes are of key importance in many fundamental cellular processes, such as growth, differentiation, and response to environmental stimuli. The study of proteins is therefore essential for the understanding of such cellular processes. However, there appear to be large gaps in our comprehensive understanding of protein function. According to a recent estimate, 75% of reported protein research focused on those 10% of known proteins that were already known when the human genome was mapped [Edwards et al., 2011]. Indeed, a considerable number of human proteins have not been tied to specific cellular functions. As of March 6, 2012, the human UniProtKB-GOA [Barrell et al., 2009] contained less than 16,000 human proteins with human-curated GO annotations. Proteomic technologies that help implicate new proteins in cellular processes are therefore of critical importance.

Although proteomic technologies have greatly evolved during the past decade [May et al., 2011], they still confront challenges, mainly sensitivity of detection. Among gel-based strategies, 2DE [O’Farrell, 1975] combined with MS for protein identification, is commonly used for the separation and quantification of thousands of proteins from complex samples, such as eukaryotic cells [Al-Ghoul et al., 2008]. Despite its potential and high resolving power, 2DE is subject to technical limitations, in particular the detection of proteins with extreme molecular weight, extreme pH, high hydrophobicity, or low abundance [Friedman et al., 2009]. To increase sensitivity, the DIGE saturation dyes [Shaw et al., 2003] were developed, requiring the use of only 5 µg of protein sample to perform a 2D-gel as Florence Pinet’s group has recently shown [Dupont et al., 2008]. Gel-free strategies are commonly based on the use of LC-MS/MS. This approach is also subject to technical limitations, rendering the detection of proteins over the whole dynamic range of protein concentrations very challenging ([Bantscheff et al., 2007]; [Bell et al., 2009]). In summary, current experimental techniques for proteomic detection and quantification, suffer from specific biases and limitations in terms of proteome coverage.
Recent computational approaches downstream of the experimental analysis place a set of identified, or differentially regulated proteins into an interpretation context with other proteins, for example in interaction networks [Aittokallio and Schwikowski, 2006]. The bioinformatic Steiner tree approach [Scott et al., 2005] starts with a set of given, experimentally determined, genes in a regulatory process. It connects the given genes into one or several regulatory subnetworks (Steiner trees) by adding protein interactions and additional genes (Steiner nodes), according to aggregate measures of reliability of protein interactions.

[Scott et al., 2005] applied this approach to a set of differentially regulated genes from Saccharomyces cerevisiae transcriptome data. They showed that the approach can identify compact subnetworks that are consistent with prior knowledge about regulatory subnetworks, and argue that the newly identified subnetworks represent plausible hypotheses for downstream analysis. [Huang and Fraenkel, 2009] extended the approach by including data from phosphoproteomic experiments and published a web server based tool called SteinerNet [Tuncbag et al., 2012] that allows users to apply a generalization of the Steiner tree approach to custom interactomes and proteins of interest. In both cases, the resulting hypotheses were demonstrated to be useful for the downstream analysis of the data in different ways, but no direct hypotheses about the relevance of the newly identified proteins and their interaction were evaluated.

Other network-based computational approaches use different principles to infer additional proteins. The MSNet approach by [Ramakrishnan et al., 2009] aims to identify additional proteins in shotgun proteomics. Candidates are those — commonly unreported — proteins for which only marginal evidence exists, e.g. proteins for which only a single peptide has been determined experimentally. A mathematical diffusion model propagates additional likelihood from high-confidence proteins along the edges of a "probabilistic functional gene network". MSNet then selects those candidate proteins that can accumulate a sufficiently high additional likelihood from their local network neighborhood. When applied to global profiling experiments, the approach was shown to significantly increase the number of identifications at a given false discovery rate. The enrichment approach by [Li et al., 2009] uses protein-protein interaction (PPI) networks to identify additional proteins that have only weak support in MS/MS data, but are members of a group of densely interacting confidently identified proteins.

All the above computational Steiner approaches aim to identify the members of functional context, but direct evidence for the actual presence of these additional proteins has been missing. With this study, we present direct experimental evidence for the presence of additional proteins inferred by a computational Steiner tree approach.
2.2. Proteomics paper

2.2.3 Materials and Methods

Overview

We used a Steiner tree-based computational approach to identify additional proteins in proteomic experiments and evaluated its utility on a dataset generated from the comparison between unaffected and affected human smooth muscle cell (SMC) protein extracts after DNase (deoxyribonuclease) I treatment. Figure 2.1 provides an illustration of the workflow. Detailed information on SMC protein extraction and 2D-DIGE analysis is provided as supporting information in the appendix (section A.1).

Steiner tree approach

The Steiner tree approach presented here builds a PPI network between proteins that differ between profiles, thereby identifying proteins missed in the experiment. It consists of the following four steps (c.f. Figure 2.2).

1. Compilation of a human PPI network
   A network of human PPIs was built by merging the two databases IntAct [Aranda et al., 2010] and biomolecular interaction network database [Bader et al., 2001]. IntAct data were retrieved in the form of a PSIMITAB-formatted file on March 29, 2010 from the IntAct website (http://www.ebi.ac.uk/intact). To acquire the biomolecular interaction network database data, the identifier search interface supplied by the Biomolecular Object Network Database webpage (http://bond.unleashedinformatics.com) was used to extract all interactions between human proteins (requiring the taxon id 9606 for both interactors). The search results were retrieved on March 29, 2010 and stored in "GI pair" format. The two networks were merged by matching protein UniProt IDs. This led to a consolidated PPI network containing 21,022 proteins and 51,975 interactions.

2. Computation of edge confidence scores
   Confidence scores were assigned to all edges of the compiled PPI network according to the functional similarity of interacting proteins. The functional similarity of two proteins was quantified using the rfunsimBP score [Schlicker and Albrecht, 2007]. rfunsimBP scores the similarity of two sets of GO biological process annotations by taking into account all pair-wise semantic similarities of terms from the two annotation sets. Its output is a similarity score, with higher scores indicating higher functional similarity. The GO annotations were retrieved from the UniProtKB-GOA project [Barrell et al., 2009] and the resulting score was transformed into edge costs by taking the inverse (cost = 1/\text{rfunsimBP}). The resulting values were restricted to a maximum of ten to avoid excessively large costs for interactions between functionally distinct proteins.

3. Determination of Steiner nodes
   The merged PPI network, costs of the edges, as well as the list of input proteins was imported into the graph tool library GOBLIN version 2.8 (http://www.math.uni-augsburg.de/~fremuth/goblin.html) see Supporting Information for source
2.2. Proteomics paper

Figure 2.1: Workflow. (1) 24 primary human SMC and one aortic SMC cultures were performed and the protein extracts treated with DNase I. (2) Using saturation DIGE labeling, 24 2D-gels were performed with the aortic SMC protein extract as standard. (3) Bioinformatic image analysis allowed the classification of protein extracts as unaffected or affected by DNase I treatment. (4) Differential spots were identified by MALDI-TOF MS. (5) The set of identified proteins was subjected to a Steiner tree-based computational approach in order to predict a network containing further proteins potentially affected by DNase I treatment. (6) After statistical evaluation of the network, (7) two Steiner node proteins were selected, (8) and validated as significantly differing by Western blot.
Figure 2.2: Steiner-tree approach. (A) Initial protein-protein interaction network including input proteins (terminal nodes: red rectangles). (B) Edge confidence score, based on functional similarity (GO biological process) of connected proteins (scale: 0 to 1). (C) Determination of the Steiner nodes (green circles) connecting all input proteins with a minimal overall sum of edge costs. Edge cost is defined as the inverse of interaction confidence. (D) Ranking of Steiner node proteins according to summed confidence scores of edges directly connecting them to terminal nodes.
2.2. Proteomics paper

An algorithm supplied by the GOBLIN library, which is based on the heuristic described in [Mehlhorn, 1988] was applied to compute a Steiner tree connecting the proteins detected in the 2D-DIGE analysis. The aim of Mehlhorn’s heuristic is to identify a tree that is able to connect all given 'terminal nodes' (i.e. the input set of proteins differing between proteome profiles) with a minimal sum of costs along its edges. The difference to the well-known minimum spanning tree approach initially described by Kruskal [Kruskal, 1956] is that a Steiner tree can include additional non-terminal nodes of the full network, so-called 'Steiner nodes' in the solution. Cytoscape v2.7 [Cline et al., 2007] was used to visualize the resulting Steiner tree. To assess the statistical significance of the obtained Steiner tree solution, we performed a simulation study in which the Steiner tree heuristic described above was executed 10,000 times on the human PPI network using as input a set of 41 target proteins randomly selected from the full PPI network. The 41 random target proteins were selected to have a similar distribution of node degrees when compared to the 41 original target proteins. In each run, we recorded the number of Steiner nodes that were used to connect the 41 terminal nodes, along with the sum of the Steiner tree edge costs. For both measures, we fitted a normal distribution to the data and computed a p-value for the probability to obtain, by chance, a network at least as small as the one determined on the experimental data.

4. Candidate selection for experimental verification

To find candidates for the experimental verification of our in-silico results, we ranked the Steiner node proteins, based on the 'augmented network' induced by the selection of all Steiner and terminal nodes and the full set of edges connecting these nodes in the compiled human PPI network. The score used for ranking the Steiner nodes was computed as the sum of the functional similarity scores of all edges that connect a given Steiner node to any of the terminal nodes. A similar ranking of the terminal node proteins, in this case summing up the scores of all edges linking a given terminal node to any other terminal node, was also performed.

Application of MSNet to the 2D-DIGE dataset

We applied the MSNet method published by [Ramakrishnan et al., 2009] to our 2D-DIGE dataset, consisting of the weighted PPI network and the set of proteins identified with different abundances between the proteome profiles of the SMC protein extracts. Since MSNet needs a protein identification probability for each protein in the network as input, we assigned a probability of 1.0 to all 41 identified proteins. Lacking identification probability scores for the remaining proteins in our weighted PPI network, we assigned them a low probability of 0.1.

We used the REST-based web API supplied by the MSNet method to upload the necessary data and tried a range of different input parameter values. In detail, we used 10, 20, 40, or 60 network reshufflings for estimation of FDRs (default value for human data: 10) and set the parameter weighing the relative contribution of the network information versus the determined MS/MS-based score to either of the values 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or 15 (default value: 10). For each parameter combination we retrieved a list of proteins with their associated MSNet identification scores, as well as score cutoffs.
2.2. Proteomics paper

corresponding to different network shuffling based significance levels represented as q-values (data downloaded on September 15, 2011).

Application of SteinerNet to the 2D-DIGE dataset

The SteinerNet method attempts to solve a generalized version of the Steiner tree problem, called prize-collecting Steiner tree (PCST). In this problem formulation, each terminal node is assigned a negative cost (prize) contribution to the overall score, and solutions to the PCST include also networks that only connect a subset of the terminal nodes.

To compare the results of our method to the SteinerNet web service, we reformatted the PPI network and the list of terminal nodes to match the input specifications of SteinerNet as stated on their webpage (http://fraenkel.mit.edu/steinernet/quickstart.html#Inputs). As protein interaction confidence, we used the functional similarity (rfunsimBP) of the interacting proteins. We increased each score of less than 0.1 to reflect the maximal edge cost of 10.0 used in our approach. As recommended on their website, the numerical prize score for all terminal node proteins was set to 1.0. The input parameter $\beta$, which controls the trade-off between including terminal nodes and excluding edges, was left at its default value of 4.0.

Upon completion of the calculations, SteinerNet outputs the resulting Steiner tree as well as an augmented network, created from all found Steiner and terminal nodes along with all interactions in the full PPI network that connect any two of those nodes. We used the same candidate-ranking scheme as described in the materials and methods section in the fourth step of the Steiner tree approach (page 25) to prioritize the Steiner nodes found SteinerNet.
2.2. Proteomics paper

2.2.4 Results

In the present study, we have used a computational Steiner tree-based approach and evaluated it on the proteomic profile modifications between two sets of human SMC protein extracts that were affected differently by DNase I treatment.

Differential protein profiles between affected and unaffected SMC protein extracts

The proteomic profile of 13 SMC cultures appeared to be affected by DNase I treatment during protein extraction, missing four intense spots that were present in the proteomic profile of unaffected SMC protein extracts.

MALDI MS was run on selected differentially expressed spots in the gel and led to identification of 41 different human proteins (for details see supplementary information on page 129 in the appendix). The 41 identified human proteins were classified into biological processes, according to GO annotations registered in SwissProt. Interestingly, more than half of the proteins whose abundance differed between unaffected and affected SMC protein extracts could be classified into two classes: proteins involved in apoptosis, and proteins involved in cell motion and actin cytoskeleton reorganization.

Proteins detected by the Steiner tree approach

To discover new proteins whose profile might systematically differ between unaffected and affected SMC protein extract, we applied a Steiner tree approach (Figure 2.2). Of the 42 terminal node proteins (bovine DNase I and 41 human proteins) supplied to the Steiner tree algorithm, one (adenylyl cyclase associated protein 1) could not be connected to the rest of the network. For the construction of a network connecting the remaining 41 proteins (Figure 2.3), 25 Steiner node proteins were selected by our algorithm. Both terminal and Steiner node proteins were ranked according to their number and confidence of interactions with other terminal nodes of the network using a functional similarity score (for details on the score used for ranking see materials and methods, page 25 and Table 2.1). Interestingly, in the PPI network, DNase I interacted solely with one protein: actin, cytoplasmic 1, which is the terminal node protein with the highest number of interactions in the Steiner network.

Statistical significance of the Steiner tree solution

To determine whether an input of 41 randomly selected — and thus largely functionally unrelated — proteins would be likely to result in a Steiner network such as then one observed in our results, we performed 10,000 simulation runs and recorded for each simulation the number of Steiner nodes used to connect the 41 terminal nodes, along with the sum of the Steiner tree edge costs. The two resulting distributions are approximately normal and clearly show that the network obtained from the proteins differing in abundance between proteome profiles was much smaller than expected when using randomly selected proteins.
2.2. Proteomics paper

<table>
<thead>
<tr>
<th>SwissProt accession</th>
<th>Gene name</th>
<th>Total interactions</th>
<th>Terminal interactions</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>P62993</td>
<td>GRB2</td>
<td>23</td>
<td>19</td>
<td>7.55</td>
</tr>
<tr>
<td>P32121</td>
<td>ARRB2</td>
<td>17</td>
<td>14</td>
<td>6.131</td>
</tr>
<tr>
<td>P49407</td>
<td>ARRB1</td>
<td>18</td>
<td>14</td>
<td>5.436</td>
</tr>
<tr>
<td>Q9Y4K3</td>
<td>TRAF6</td>
<td>15</td>
<td>10</td>
<td>4.705</td>
</tr>
<tr>
<td>P16104</td>
<td>H2AFX</td>
<td>16</td>
<td>12</td>
<td>4.387</td>
</tr>
<tr>
<td>P00533</td>
<td>EGFR</td>
<td>10</td>
<td>8</td>
<td>3.756</td>
</tr>
<tr>
<td>Q9NYJ8</td>
<td>MAP3K7IP2</td>
<td>8</td>
<td>6</td>
<td>2.968</td>
</tr>
<tr>
<td>P62988</td>
<td>UBC</td>
<td>10</td>
<td>3</td>
<td>2.267</td>
</tr>
<tr>
<td>P23528</td>
<td>CFL1</td>
<td>7</td>
<td>3</td>
<td>2.045</td>
</tr>
<tr>
<td>P40337</td>
<td>VHL</td>
<td>6</td>
<td>5</td>
<td>1.958</td>
</tr>
<tr>
<td>P05067</td>
<td>APP</td>
<td>4</td>
<td>4</td>
<td>1.818</td>
</tr>
<tr>
<td>P01106</td>
<td>MYC</td>
<td>4</td>
<td>3</td>
<td>1.614</td>
</tr>
<tr>
<td>P07996</td>
<td>THBS1</td>
<td>2</td>
<td>2</td>
<td>1.542</td>
</tr>
<tr>
<td>Q9NZJ5</td>
<td>EIF2AK3</td>
<td>2</td>
<td>2</td>
<td>1.27</td>
</tr>
<tr>
<td>Q9NP98</td>
<td>MYOZ1</td>
<td>2</td>
<td>2</td>
<td>1.255</td>
</tr>
<tr>
<td>Q92574</td>
<td>TSC1</td>
<td>3</td>
<td>3</td>
<td>1.228</td>
</tr>
<tr>
<td>P23508</td>
<td>MCC</td>
<td>8</td>
<td>4</td>
<td>1.052</td>
</tr>
<tr>
<td>Q7L5N1</td>
<td>COPS6</td>
<td>3</td>
<td>1</td>
<td>0.997</td>
</tr>
<tr>
<td>Q99816</td>
<td>TSG101</td>
<td>3</td>
<td>1</td>
<td>0.757</td>
</tr>
<tr>
<td>P04637</td>
<td>TP53</td>
<td>3</td>
<td>1</td>
<td>0.657</td>
</tr>
<tr>
<td>Q8IZH2</td>
<td>XRN1</td>
<td>3</td>
<td>1</td>
<td>0.583</td>
</tr>
<tr>
<td>P12004</td>
<td>PCNA</td>
<td>3</td>
<td>1</td>
<td>0.426</td>
</tr>
<tr>
<td>P60953</td>
<td>CDC42</td>
<td>3</td>
<td>1</td>
<td>0.415</td>
</tr>
<tr>
<td>Q9H0D6</td>
<td>XRN2</td>
<td>2</td>
<td>1</td>
<td>0.379</td>
</tr>
<tr>
<td>Q01780</td>
<td>EXOSC10</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2.1: Proteins included in the Steiner tree solution
2.2. Proteomics paper

Figure 2.3: Computed Steiner tree. Terminal nodes corresponding to human proteins are represented as red rectangles, DNase I as a blue diamond. Steiner node proteins are represented as green circles. Edge gray levels indicate confidence scores. Interactions tested in human samples are represented by solid lines and interactions tested only in other organisms are shown as dashed lines.
2.2. Proteomics paper

proteins with a comparable number of connections in the PPI network (supplementary figure A.1). In fact, none of the 10,000 simulations yielded a network with scores as low as the ones obtained from the experimental data with respect to either of the measures. The corresponding p-values indicated statistical significance for both the number of Steiner nodes \((3.6 \times 10^{-5})\) and the sum of edge costs \((2.3 \times 10^{-5})\), suggesting overall close relatedness of the input proteins in the PPI network and corroborating the hypothesis that the proteins whose abundance differed between profiles were indeed affected by the DNase I treatment.

Biological validation of two predicted proteins

To biologically validate some of the proteins the Steiner tree approach predicted to be affected by the DNase I treatment, we examined the second and third protein in the candidate list, beta-arrestin 1 and beta-arrestin 2 (Table 2.1), as these proteins also showed an interaction with actin, cytoplasmic 1. Moreover, both beta-arrestins together interacted with 17 of the 41 terminal node proteins (Figure 2.4A).

Quantification of beta-arrestin 1 and beta-arrestin 2 was performed on unaffected (n = 10) and affected (n = 11) SMC protein extracts by Western blot (Figure 2.4B). Western blot analysis was performed according to a protocol detailed in the appendix on page 129. The results clearly showed that protein abundance of both beta-arrestins were higher in unaffected compared to affected samples.

MSNet results on the 2D-DIGE dataset

In order to evaluate whether the MSNet method [Ramakrishnan et al., 2009] would be able to identify the same proteins predicted in the present study, we applied MSNet to our 2D-DIGE data set, e.g. the weighted PPI network and the set of 41 experimentally identified proteins. In the input for the MSNet method, we set a high identification probability to proteins found in the 2D-DIGE data set and a basal low probability to all other proteins in the network, thus allowing the method to identify any protein reachable from the set of experimentally detected proteins.

Using a range of different input parameter settings (see materials and methods, page 25), MSNet was unable to predict the two beta-arrestins validated by Western blots at a reasonable q-value cutoff. Regardless of the supplied input parameters, the MSNet probability score for the two beta-arrestins never exceeded 0.12. Selecting a prediction score cutoff low enough to identify the beta-arrestins led on average to the prediction of 2356 out of the roughly 21,000 proteins in the network, which is also reflected by the low significance level estimated at a q-value of 0.3.

SteinerNet results on the 2D-DIGE dataset

Using a generalization of the Steiner tree problem, the web server SteinerNet [Tuncbag et al., 2012] also computes Steiner trees given a PPI network and a set of terminal node proteins (cf. materials and methods, page 26). Running SteinerNet on our scored PPI network and the 41 experimentally identified proteins resulted in a list of 74 Steiner
Figure 2.4: Protein-protein interactions and experimental profile of beta-arrestins. (A) The network shows all interactions between the two beta-arrestins and the set of terminal and Steiner node proteins. Terminal node proteins are represented as red rectangles, Steiner node proteins as green circles, and beta-arrestins (ARRB1 and ARRB2) as yellow diamonds. Edge gray levels indicate confidence scores. Interactions tested in human samples are represented by solid lines and interactions tested only in other organisms are shown as dashed lines. (B) Beta-arrestin 1 and beta-arrestin 2 were quantified by Western blot analysis of 50 µg of each protein from extracts of unaffected (n = 10) and affected (n = 11) SMC. Western blots from three representative samples for each group and a histogram representing total amounts over all samples are shown.
proteins that connected all 41 input proteins with a total cost of 174.7 (compared to 25 Steiner proteins connecting all inputs with total cost of 140.5 in our solution). Since SteinerNet attempts to solve a generalization of the Steiner tree problem and finds a solution that is also valid for the regular Steiner tree problem (all terminal nodes are connected), the higher cost of its solution relative to our implementation can only be attributed to the different heuristic it uses. The list of 74 Steiner proteins found by SteinerNet overlapped the list of proteins found by our approach (see Supplementary table 4 in section A.1 of the appendix), but was considerably longer. While 18 of the 25 proteins identified by our method were also contained in the SteinerNet result, some high-ranking proteins were missed (e.g. beta-arrestin-2 and EGFR). At the same time, the SteinerNet solution introduced few high-scoring new candidates. Only one protein, TRAF2, was not found by our method and ranked in the top ten, when the two result lists were combined. The average score of the top ten candidates in the SteinerNet list was 3.53, while our method achieved an average score of 4.12 within its top ten candidates. These results indicate that our method not only found a better solution of the Steiner tree problem, but also produced a protein candidate ranking in which high-ranking proteins had more interactions with — and were functionally more similar to — the input proteins.
2.2.5 Discussion

The aim of the present study was to probe the potential of the computational Steiner tree approach to identify novel proteins that are not detected by a differential proteomics approach (saturation DIGE analysis). To our knowledge, this is the first experimental validation of this approach, in its first reported application to only a single proteomic dataset, and to samples of a complexity higher than yeast cell lysate. The present paper presents clear evidence that the combined experimental-computational approach gives access to proteins that were not found by 2D-DIGE analysis alone.

The Steiner tree approach is a straightforward and widely applicable method to implicate additional proteins in differential proteomics

The Steiner tree approach, in its different variations, is a computational procedure that requires, besides an experimentally determined set of proteins, only access to public databases on protein function, protein interactions, and relatively mild computational means. Our study shows that it can be successfully applied even to complex human samples. In the future, protein interaction and functional knowledge databases can be expected to become more complete and reliable, and the computational means will only improve as well. This suggests that the Steiner tree approach represents an inexpensive, widely applicable technique that can be routinely applied to implicate additional, potentially undetected proteins in differential proteomics experiments, such as 2DE or even LC-MS/MS.

A few recent studies have previously proposed Steiner tree techniques to integrate experimental evidence into functional subnetworks, but our study suggests a simpler, more direct, and experimentally testable use of the technique. We applied it here to a single proteomic dataset using only public interactome and GO annotation data, whereas previous applications ([Scott et al., 2005]; [Huang and Fraenkel, 2009]) require additional quantitative data such as phosphoproteomic, ChIP-chip, and transcriptomic measurements. Previous work focused on elucidating regulatory subnetworks or pathways potentially implicated in the process under study, rendering the downstream validation and use of the results less obvious. The simple and straightforward protein detection and ranking scheme presented here allows direct experimental testing of the predicted proteins. Moreover, the previously published methods were tailored to data obtained in Saccharomyces cerevisiae, which is to be attributed to the wealth of knowledge and high-throughput experimental results collected for this model organism. By applying our method successfully to a set of experimentally detected human proteins, we were able to show that the human interactome now approaches a level of completeness that enables the application of such graph-based prediction methods.

A key difference to the MSNet method is that a diffusion model tends to select additional proteins whenever they are close in the network to many input proteins. A Steiner tree represents a maximally sparse connecting structure between input proteins, which can be justified using an obvious parsimony argument. We deem the former approach more suitable for protein identification in global profiling experiments, whereas the Steiner tree approach, which assumes that the differentially regulated proteins tend to form a connected subnetwork, may be most suitable for data from differential proteomics. The fact
2.2. Proteomics paper

that we were not able to predict the two additional proteins identified by our method using the MSNet method, supports the idea that the two approaches may indeed have different application domains: the Steiner tree approach may be more suitable for "network-local" datasets typical for differential proteomics, whereas diffusion-based approaches may be better in the case of global datasets. The enrichment approach by \[Li et al., 2009\] can also be viewed as complementary to the Steiner approach, as it aims at complementing (typically already strongly connected) sets of input proteins by additional proteins in the same biological complex.

Although we could only test two of the 25 predicted proteins due to the limited amount of protein extract gathered from the SMC samples, the fact that we were able to clearly confirm those two proteins (beta-arrestins) experimentally still appears significant, given that the approach uses no knowledge that could have specifically favored the detection of these proteins. An explanation for the success of our prediction strategy may lie in the fact that it favors proteins that are interacting with many of the experimentally detected proteins in the PPI network. It can thus be understood as an extension of the "guilt-by-association" rule, which is commonly used in protein function classification and prediction of disease-causing genes (\[Wang and Marcotte, 2010\]; \[Schwikowski et al., 2000\]).

Sensitivity of saturation DIGE versus the Steiner tree approach

The effect of DNase I on the novel proteins added by the Steiner tree approach was assessed by measuring levels of abundance of two confidently predicted Steiner node proteins, beta-arrestin 1 and beta-arrestin 2. These proteins showed such a strong differential abundance in SMC protein extracts that beta-arrestin 1 and beta-arrestin 2 were not at all detected in unaffected samples (figure 2.4B).

While the beta-arrestins were detected by the Steiner tree approach, their differential profile passed unnoticed by the experimental proteomic approach. Their biochemical properties (supplementary figure A.2) do not correspond to known biases or specific limitations of saturation DIGE technology. We verified that the bioinformatic analysis of DIGE images carried out by the SameSpots software should be able to detect and quantify differences in abundance levels of a magnitude as observed in the Western blots for the beta-arrestins. We also verified that the expected Mr and pI of the two beta-arrestins after DIGE labeling did not correspond to any of the spots that failed protein identification by MALDI-TOF MS. We therefore suspect that the beta-arrestins remained undetected in the experimental proteomic approach due to their low abundance in the SMC extracts. We note that, also for LC-MS/MS-based approaches, each beta-arrestin would have likely generated only a single peak, as these proteins were undetectable even by Western blot in one set of SMC extracts — and would therefore not have been detected with different abundance levels by most standard data analysis approaches.

However, our and any other Steiner tree-based methods suffer from other biases and limitations. One of them could be called "region bias". As the Steiner tree approach depends on connecting previously detected proteins in an experiment, it is unable to penetrate into regions of the PPI network in which no proteins have been detected experimentally. Conversely, the uneven quality and coverage in current protein interaction databases \[Ramírez et al., 2007\] can cause biases against, or entirely prevent, the detection of certain proteins,
2.2. Proteomics paper

for instance, those that have not yet been tested for interactions, which may also imply a slight bias against low abundance [Ivanic et al., 2009]. Similarly, as the edge weights of our method are biased toward including interactions with similar GO annotations, it is biased against unannotated proteins. Finally, other interactions than PPIs may be involved in the network propagation of molecular changes, and may have to be included in computational methods such as the one presented here.

We therefore consider the Steiner tree approach not an endpoint, but—akin to other bioinformatics methods—rather a tool for the generation and prioritization of hypotheses that can be experimentally validated by low- and medium-throughput experimental methods. The reason for the apparent effectiveness of the Steiner method may lie in the independence of the above biases relative to the biases of the experimental methods, a principle that may be key for the discovery of other large-scale data integration approaches as well.

Conclusions

When applied to proteins found by 2D-DIGE analysis, the computational Steiner tree approach we present here identified additional proteins that were subsequently validated using Western blots. The biochemical properties of the additional proteins did not correspond to the specific limitations of the 2D-DIGE platform, suggesting that the Steiner tree approach may be a simple, cost-effective, and generic computational tool that can also be useful on other platforms, such as LC-MS/MS, for differential proteomics experiments.

While network-based approaches are not free of their own limitations, their potential to predict significant numbers of additional proteins suggests that they might also be followed by highly sensitive, targeted, medium-throughput MS techniques such as SRM. In the future, such strategies will be able to draw on increasingly complete and accurate interaction networks. Network-based approaches might therefore prove to be important new tools to extend the reach of purely experimental methods.
2.2. Proteomics paper
Chapter 3

Plant responses to environmental stresses

3.1 Background

This project was conceived and carried out in collaboration with Pr. Heribert Hirt. Pr. Hirt leads the "Unité de Recherche en Génomique Végétale" (URGV) in Evry. His group’s research goal is to understand the molecular mechanisms of how plants sense, transduce and adapt to changes in environmental conditions. In order to achieve this goal, they use transcriptome, proteome, phosphoproteome and metabolome profiling technologies as well as phenotyping and screening of stress signaling mutants. The work done in this project was focused on analyzing transcriptomic measurements of time-resolved responses to abiotic stresses with modeling techniques to identify regulatory networks involved in stress signaling and response.

3.1.1 Sensing of and adaptation to environmental stresses

Plants, as sessile organisms, heavily rely on detection and adaptation mechanisms to environmental perturbations and presence of pathogens. They have thus developed robust and intricate molecular machinery to sense changes in environmental conditions and can adapt to a variety of environmental threats by inducing specific sets of gene products and metabolites. The molecular machinery allowing these adaptations is very complex and includes different stress sensing mechanisms, signal transduction pathways which are interlinked through cross-talk and downstream effectors that lead to changes in plant metabolism and targeted expression of proteins to counteract the perceived stress. One example of an extensively studied plant stress sensing and response system — with a defined sensor, well-characterized signal transduction components and downstream stress response products — is the response to the presence of bacteria exhibiting flagella. Flagellin, a protein component of flagella, a bacterial motile apparatus, is sensed by Arabidopsis with the aid of the surface-localized receptor FLAGELLIN SENSING 2 (FLS2) [Gómez-Gómez and Boller, 2002]. FLS2 was shown to rapidly form a complex with the brassinosteroid-associated kinase 1 (BAK1), a receptor-like kinase previously implicated in hormone sig-
3.1. Background

Signaling, upon stimulation with flagellin ([Heese et al., 2007], [Chinchilla et al., 2007]). The FLS2-BAK1 complex then leads to activation of mitogen-activated protein kinase (MAPK) cascades [Colcombet and Hirt, 2008]. MAPK cascades are composed of three layers of protein kinases (consisting of MAPKKKs, MAPKKs and MAPKs respectively), with each layer in turn activating kinases of the following layer by phosphorylating specific sites in their activation loops. The final layer of this phospho-relay system is composed of MAPKs some of whose activation leads to initiation of defense responses against pathogens through activation of specific downstream transcription factors and production of plant hormones (like e.g. ethylene [Liu and Zhang, 2004]). In Arabidopsis thaliana 110 MAPK-related kinases are known, out of which 80 are put in the first, 10 in the second and 20 in the third layer of the cascades according to similarity with well studied MAPK phosho-relay systems in animals and fungi [MAPK-Group, 2002]. This very large number of MAPK-related kinases along with their specificity for phosphorylation targets and modes of activation builds the basis of a very complex signal transduction system capable of integrating different stimuli through cross-talk between individual MAPK-modules. Consequently, MAPK cascades have been implicated not only in the response to bacteria but also in developmental processes (e.g. epidermal patterning and embryo development), in the response to various plant hormones (e.g. ethylene, abscisic acid and jasmonic acid) and to a number of abiotic stresses (e.g. cold and salt [Teige et al., 2004], wounding, heat, UV light and osmotic shock [Droillard et al., 2004]).

Temperature stresses (cold or heat), water stresses (flooding or dessication) and high salinity are the abiotic stresses with the highest impact on agriculture by causing, potentially catastrophic, crop loss. In the following, we will briefly summarize what is known about the above mentioned stresses’ mode of action and how plants are sensing and adapting to them.

Cold stress

Low temperature affects water and nutrient uptake, protein and nucleic acid conformation and membrane fluidity [Chinnusamy et al., 2007]. Furthermore, it influences cellular metabolism by reducing the rates of biochemical reactions. Two types of low temperature stress have to be distinguished: Freezing, i.e. exposure to temperatures below 0°C, and chilling, a more moderate low temperature stress (< 20°C). Freezing stress induces ice formation in intercellular spaces and results in a drop in water potential outside the cell. This, in turn, leads to movement of unfrozen water from inside the cell to the intercellular spaces leading to dehydration of the plant and high osmolarity of the intracellular fluid [Thomashow, 1999]. Freezing, but not — or to lesser degrees — chilling stress thus induces dehydration and osmotic stresses. Many plants are capable of withstanding freezing stress if they have been subjected to chilling stress before. This sensing of low temperatures and development of freezing tolerance is known as cold acclimation and includes pro-active metabolic and gene-expression changes preparing the plant for freezing effects. It can thus be argued that, while little to no osmotic and dehydration stress occur at chilling temperatures, the chilling response should include some elements of osmotic and dehydration stress responses.

In comparison to the sensing of bacterial products like flagellin, the exact sensing mechanism of most abiotic stresses is less well characterized. There is, however, strong ev-
3.1. Background

Evidence implicating the increased rigidity of the plasma membrane at low temperatures as a cold sensing mechanism [Vaultier et al., 2006]. Furthermore, cytosolic Ca\(^{2+}\) levels increase upon cold shock, possibly mediated through membrane rigidification-activated Ca\(^{2+}\) channels. The calcium signal can then be amplified and transduced e.g. by calcium-dependent protein kinases [Komatsu et al., 2007]. There is evidence that oscillations in calcium concentrations can be perceived distinctly in different tissue types and lead to initiation of responses to a number of different stresses ([Kiegle et al., 2000]; [Mehlmer et al., 2010]; [Wurzinger et al., 2011]).

A well studied signaling component of cold acclimation is the ICE1-CBF transcriptional cascade. C-repeat binding factors (CBFs) are transcription factors induced within 15 minutes of cold stress exposure and have been shown to bind to a motif termed C-repeat (CRT)/drought-responsive element (DRE), present in the promoters of many cold-induced genes [Liu et al., 1998]. Due to their early induction and numerous cold-induced targets, CBFs have been termed "master-switches" of cold acclimation. More recently, the upstream effector of CBFs, inducer of CBF expression 1 (ICE1), a constitutively expressed transcription factor [Chinnusamy et al., 2003] whose activity is under post-translational control — through a competition between stabilizing sumoylation by SIZ1 [Miura et al., 2007] and ubiquitination-mediated degradation by HOS1 [Dong et al., 2006] — has been identified. The exact connection between cold-sensing and the ICE1-CBF regulon, however, still remains elusive [Miura, 2013]. The targets of the CBF transcription factors are cold-induced genes, many out of which encode hydrophilic peptides (e.g. from the family of Late Embryogeneis Abundant (LEA) proteins) that are thought to maintain membrane fluidity [Thomashow, 1999]. Furthermore, transcription factors RAP2.1 (from the DREBA5 family) and RAP2.6 (from the ERF family) are induced by CBFs and presumably control subregulons of the CBF regulon [Fowler and Thomashow, 2002].

Drought and osmotic stress

Dehydration brought forward by long periods of drought has some molecular effects overlapping with freezing stress. The membrane fluidity is altered — in severe cases leading to disruption of the bilayer structure and loss of cellular compartmentalization — and osmolarity is increased due to loss of intracellular water. Consequently, plants subjected to drought and freezing stress both also face an osmotic stress component.

On a systemic level, however, drought responses are quite different from cold responses. The main focus is put on minimizing water loss by e.g. reducing photosynthesis, which consumes water, closing stomata to minimize evaporation, halting growth and activating respiration. One of the earliest discovered drought sensing mechanisms was published by [Urao et al., 1999] who showed that water deficit, followed by the resulting osmotic stress, triggered the expression of a putative osmosensor AtHK1, a transmembrane protein with a histidine kinase domain. Moreover, they showed that AtHK1 was able to rescue the reactivity to hyperosmolarity of yeast cells depleted of the yeast osmosensor SLN1. In yeast, SLN1 activates the MAPK high-osmolarity glycerol response 1 (HOG1), suggesting that AtHK1 might transduce its signal via a MAPK cascade as well. In later work ([Reiser et al., 2003]; [Tran et al., 2007]), three more histidine kinases involved in osmosensing were identified and shown to be involved in regulation of drought, osmotic and salt stress responses as well as in abscisic acid signaling.
3.1. Background

The phytohormone abscisic acid (ABA) is a plant growth regulator, imposing dormancy in developing seeds. Cellular ABA levels have been shown to increase rapidly upon drought stress and have been implicated in systemic signaling leading to stomatal closure as well as in induction of drought and salinity stress response genes [Shinozaki and Yamaguchi-Shinozaki, 2006]. Drought responses can be broadly classified into two pathways: one is acting independent of ABA, the activation of the other one is dependent on ABA [Yamaguchi-Shinozaki and Shinozaki, 2005].

The best-studied ABA-independent drought response pathway involves the DREB2 (DRE-binding) family of transcription factors. The family consists of eight members termed DREB2A to DREB2H. DREB2A and DREB2B are induced under drought and high salinity but not under cold conditions [Liu et al., 1998] and bind the same CRT/DRE core motif as the cold-inducible CBFs. Unlike CBFs, overexpression of DREB2 genes did not result in any phenotypic changes. This suggests a post-transcriptional regulation of DREB2 genes, which was later identified to be mediated by a regulatory region of the protein, whose removal resulted in a constitutively active form. As CBFs and DREB2 proteins bind the same CRT/DRE core motif, the targets of the cold- and drought/osmotic-induced regulons show significant overlap. The canonical downstream target of the DREB2 transcription factors, RD(Responsive to Dehydration)29A/LTI(Low Temperature Induced)78 for example has been shown to be induced by cold, drought as well as high salinity stresses [Yamaguchi-Shinozaki and Shinozaki, 1994].

Another ABA-independent pathway responding to drought and high salinity stress is regulated through a set of NAC [for NAM (no apical meristem), ATAF, CUC (cup-shaped cotyledon)] transcription factors. This NAC-regulon was mainly studied using the Early Responsive to Dehydration 1 (ERD1) gene as a read-out. The promoter of ERD1 contains two cis-acting binding motifs. ANAC19, ANAC055 and ANAC072 were shown to bind to one of those motifs in the promoter of ERD1 and, when overexpressed, induced expression of a number of drought response genes but failed to induce ERD1 expression [Tran et al., 2004]. Later, a zinc finger homeodomain transcription factor ZFHD1 was shown to bind to the other cis-element in the promoter of ERD1 and to lead, when overexpressed along with the NAC factors, to induction of ERD1 expression [Tran et al., 2006].

Cellular ABA levels are under a dynamic balance mediated by ABA biosynthesis, catabolism and transport. ABA can be inactivated by conjugation with glucose to form ABA glucosyl ester (ABA-GE). ABA-GE is stored in vacuoles and the apoplast [Cutler and Krochko, 1999]. Under dehydration conditions, inactivated ABA-GE can be reactivated by β-glucosidases BG1 and BG2, whose expression has been shown to be induced by drought and salt stress [Xu et al., 2012]. This reactivation of ABA allows for a very fast stress response as it does not depend on de-novo biosynthesis of ABA. The signaling events elicited by ABA stimulation have recently been clarified by the discovery of a central signaling module (c.f. the review by Danquah et al., 2013b). It is made up of three classes of proteins: the ABA receptors are from the class of PYR/PYL/RCAR proteins. Protein Phosphatase 2Cs (PP2Cs) act as negative regulators and SNF1-related protein kinases 2 (SNRK2s) are positive regulators of ABA signaling. In the presence of ABA, PP2Cs form a complex with the ABA receptors, blocking their phosphatase activity. This, in turn, allows activation of SNRK2s, which are inactivated (dephosphorylated).
3.1. Background

by PP2Cs without the presence of ABA. The SNRK2s then continue to phosphorylate and thus activate downstream transcription factors termed ABRE-binding factors (ABFs) which bind to ABA-responsive element (ABRE) motifs in target gene promoters. Among this class of transcription factors, ABF2, ABF3 and ABF4 have been most prominently shown to be functioning downstream of ABA signaling in response to drought and high salinity [Fujita et al., 2005].

Salt stress

High salinity in the soil leads, as indicated above, to an osmotic gradient drawing water out of the plant’s roots. This osmotic stress leading to dehydration is a common element of drought and salt stress and is perceived by the plant in a similar way. High salinity does, however, lead to an additional stress, namely ion toxicity. Ion toxicity is the result of replacement of K\(^+\) by Na\(^+\) in biochemical reactions, and Na\(^+\) and Cl\(^-\)-induced conformational changes of proteins [Chinnusamy et al., 2006]. Plants cope with excess Na\(^+\) concentrations by actively excreting it or transporting it to structures less susceptible to ion toxicity. The pathway of salt overly sensitive (SOS) proteins was discovered using forward genetic screening for salt sensitive mutants ([Wu et al., 1996]; [Zhu et al., 1998]). It is composed of three proteins (SOS1, SOS2, SOS3) instrumental in conferring salt tolerance of Arabidopsis thaliana. SOS1 encodes a plasma-membrane localized Na\(^+\)/H\(^+\) antiporter, SOS2 is a protein kinase and SOS3 is a Ca\(^{2+}\)-binding protein. The current model of the SOS signaling pathway states that SOS3, stimulated by elevated Ca\(^{2+}\) concentrations, binds to SOS2, thus forming a complex needed to activate SOS2’s kinase function. The SOS2/SOS3 complex is recruited to the membrane and increases SOS1’s Na\(^+\) transporting function by phosphorylating it (see review by [Zhang and Shi, 2013]). In addition to the SOS pathway, two other proteins have been shown to play a pivotal role in salt tolerance: the Na\(^+\) transporter AtHKT1 is thought to facilitate transport of sodium from the shoot to the root tissue [Rus et al., 2001], while the function of the Na\(^+\)/H\(^+\) antiporter NHX1 is in compartmentation of Na\(^+\) into vacuoles [Gaxiola et al., 1999].

Heat stress

Exposure of plants to elevated temperatures, heat stress, can lead to severe cellular damage resulting in cell death. The maximal temperature tolerated by plants is highly dependent on the ambient temperature the plant is used to, whether the plant has perceived elevated temperatures before (heat acclimation) and the duration of heat exposure. Rapid, direct effects of heat include protein denaturation and aggregation as well as increased fluidity of lipids in membranes. Slower, indirect effects are inactivation of enzymes in chloroplasts and mitochondria, blocking of protein biosynthesis as well as membrane and protein degradation [Schöffl et al., 1999]. The deleterious effects of high temperatures on cellular integrity, metabolism and protein homeostasis eventually lead to starvation, growth inhibition, production of reactive oxygen species (ROS) and toxic compounds. On a systemic level, heat adaptation of plants is mediated by opening of stomata to cool down the leaves through increased evaporation, synthesis of chaperone proteins stabilizing protein folding, activation of damage repair mechanisms and detoxification.
3.1. Background

Heat shock proteins (HSPs) are constitutively expressed molecular chaperones implicated in protein folding, quality control and maintenance of membrane integrity [Nakamoto and Vigh, 2007]. As such, they play a pivotal role in the heat stress response system, as shown by their conservation throughout prokaryotes and eukaryotes [Scharf et al., 2012]. Sensing of heat stress is thought to be mediated by recognition of deviation from protein homeostasis, possibly through recognizing the depletion of free HSPs. Upon stress perception, heat shock factors (HSFs), a class of transcription factors either activated by oligomerization and localization to the nucleus or newly produced during stress, lead to the induction of heat shock responsive (hs) genes containing heat shock element (HSE) motifs in their promoters. HSPs are among the hs genes and are newly synthesized to replenish the HSP pool [Baniwal et al., 2004]. There is good evidence that some of the HSPs are involved in the restoration of the inactive state of HSFs in a negative feedback loop from studies in human cells [Guo et al., 2001], Arabidopsis [Kim and Schöffl, 2002] and tomato [Port et al., 2004].

In Arabidopsis, 21 HSF genes have been identified. They can be classified into class A (15 members), class B (5 members) and class C (1 member) according to structural similarities. In tomato, the protein HSFA1a is constitutively expressed and has been established as a master regulator of heat shock response in knock-down experiments [Mishra et al., 2002]. In Arabidopsis, the four members of the HSFA1 group are hypothesized to function similar to the single protein in the tomato HSFA1 group. Redundancy in terms of their function can be inferred from the fact that a quadruple knock-out of all HSFA1s was needed to show increased heat sensitivity and reduced hs gene induction [Liu et al., 2011]. Transcriptomic analyses of single, double, and triple knock-out plants do, however, indicate a certain level of differential target specificity of the four HSFA1s. HSFA2, structurally similar to HSFA1, is only expressed in stressed plants. In tomato, HSFA1 and HSFA2 have been shown to form heterodimers, which are much more efficient in inducing hs gene induction than either of the factors alone. Stress tolerance/sensitivity of Arabidopsis HSFA2 overexpression/knock-out plants indicate a major role for HSFA2 not only in heat, but also in salt and oxidative stress. Congruent with this, genes encoding for galactinol synthases (GolS), involved in the synthesis of osmolytes, and ascorbate peroxidases (APX), which are involved in ROS scavenging, are highly induced in heat stress and their induction has been shown to be dependent on HSFA2 [Nishizawa et al., 2006]. Arabidopsis HSFA3 expression is equally induced not only under heat, but also under drought stress conditions and it has been shown that DREB2A and DREB2B are able to induce its expression using a reporter construct containing the HSFA3 promoter in fusion to GUS [Schramm et al., 2007]. HSFA6a and HSFA6b are highly induced under salt and drought stress and their induction has been shown to be reduced in plants deficient in ABA-dependent drought signaling [Yoshida et al., 2010]. HSFB genes are thought to be transcriptional repressors of hs genes needed in the recovery from a perceived heat stress [Scharf et al., 2012]. The function and target specificity of the only HSFC gene HSFC1 has not yet been described in more detail.

Despite the above mentioned findings, the exact signaling mechanisms upstream of HSFs as well as the mechanisms behind their differential target specificities are still poorly understood. Recently, the NAC transcription factor NAC019, implicated in the ABA-independent drought response, has been found to bind to the promoters of several HSFs [Guan et al., 2014]. The authors furthermore show that expression of most HSFs is up-
regulated in overexpressors of NAC019 and postulate that NAC019 is activated through dephosphorylation by the protein phosphatase RCF2.

**Stress cross-talk**

As shown in the previous sections, there is considerable cross-talk between signaling of and overlap of responses to different abiotic stresses. This is, in part, due to very central and ubiquitously used signaling components like the MAPK cascades and calcium signaling as a response to membrane fluidity changes. Additionally, a number of stresses lead to similar cellular problems like e.g. the production of ROS and the optimal way to react to a present stress depends on the plants’ developmental and physiological status. Consequently, the cellular machinery for sensing of and response to different stresses is highly complex and integrates various signals external to the actual stress, like e.g. plant hormone and metabolite levels. Due to this high level of complexity, to understand how exactly plants respond to abiotic stresses is still a very hard challenge.

### 3.1.2 Effect of stress combinations

Despite their remarkable potential for adaptation, frequently co-occurring stresses, like long periods of drought and simultaneously elevated temperatures, often become too harsh for plants to survive or yield crop. In the case of heat and drought stress, this can be ascribed to contradicting adaptation strategies to the two stresses. During heat stress, plants open their stomata to cool their leaves by transpiration. However, if heat stress is combined with drought, plants are not able to open their stomata and their leaf temperature is higher [Rizhsky, 2002]. Figure 3.1 shows that a combination of heat and drought stresses has a more detrimental effect on vegetation health than drought stress occurring without elevated temperatures.

Despite these insights, most plant abiotic stress studies have been performed on single stresses and comparably little genome-scale information is available on the mechanisms and effects of stress combinations [Mittler, 2006]. For heat and drought combinations, a previous study, [Rizhsky, 2004], showed a surprisingly low overlap of induced genes between heat and drought applied separately and a combination of heat and drought stress, indicating that the adaptation to a combined stress requires more than just switching on both stress response systems. Unfortunately, the study by Rizhsky et al. measured the plants’ responses to stresses only at a single time-point, which makes it hard to infer the key regulatory events occurring during stress exposure.

### 3.1.3 Available genome-scale data

Comparison and joint analysis of genome-scale transcriptomic data sets measuring plant responses to abiotic stresses has been hampered by inconsistent data due to technical differences (e.g. different array designs used), differences in the modes of stress application (e.g. exact stress parameters, duration of stress exposure and the time after stress onset the samples were taken), differences in experimental conditions (like e.g. light intensity)
3.1. Background

Figure 3.1: The effect of a combination of heat and drought stress on the US vegetation health: During August 1997 (1), no prolonged heat or drought period was observed. In August 2000 (2), heat and drought periods coincided leading to similar levels of dry and wet areas as in August 2002 (3), during which a drought but no heat wave was recorded. The maps on the right clearly show that coinciding heat and drought conditions have far stronger influences on vegetation health than drought alone. This figure was copied from Figure 1 b) in Mittler, 2006

and differences in plant material used (e.g. age/developmental status, type of plant tissue sampled). Additionally, very few genome-scale studies (e.g. Fowler and Thomashow, 2002, Seki et al., 2002) measured significant numbers of time-points during stress exposure and/or during recovery from stress.

AtGenExpress

To overcome these limitations, the Arabidopsis Functional Genomics Network (AFGN) coordinated the AtGenExpress project [Kilian et al., 2007], a comprehensive Arabidopsis thaliana genome transcript expression study performed using the Affymetrix ATH1 microarray, which measures the expression of around 24,000 protein-coding genes. In this project, considerable care was taken to define strict experimental protocols and time-resolved transcriptomic responses to heat, cold, drought, salt, high osmolarity, UV-B light and wounding stresses were measured along with a control time-series of unstressed plants. To facilitate stress application and harvesting, Kilian et al. decided to initially cultivate the plants for 13 days at 24°C under sterile conditions on polypropylene rafts in growth boxes containing liquid medium (see Figure 3.2). The boxes were closed with a lid con-
3.1. Background

Taining an opening for air ventilation. Thirteen days after sowing, the plant-containing rafts were transferred to new growth boxes containing fresh medium and were cultivated for five additional days before initiating stress treatments [Kilian et al., 2007]. Of the seven stresses, three — cold, osmotic and salt stress — were applied in a permanent manner with the stress exposure lasting over the full 24h measurement period. The remaining four stresses — heat, drought, UV-B and wounding stress — were applied in a transient manner with measurements being taken during (or immediately after) stress exposure as well as after return of the plants to normal growth conditions to capture the recovery process. At each measured time-point, a pool of root and shoot tissues of nine plants was used for transcriptomic analysis and two replicate experiments were performed for each stress. The modes of application for the stresses we focused our analysis on (cold, osmotic and salt as permanently applied stresses and the transiently applied heat and drought stresses) will be described in more detail in the following:

- **Cold stress** was applied by putting the boxes on ice for rapid cooling and keeping them in a cold room at $4^\circ\text{C}$ until harvest.

- **Osmotic and salt stresses** were applied by removing the plant-containing rafts from the growth boxes and adding mannitol or NaCl to the medium to obtain a final concentration of 300 mM and 150 mM, respectively. The rafts were then returned to the boxes and the boxes kept under standard growth conditions until harvest.

- **Heat stress** was applied by putting the growth boxes in an incubator set to $38^\circ\text{C}$, keeping them there for 3 hours and subsequently returning them to standard growth conditions in a phytochamber until harvest.

- **Drought stress** was applied by removing the rafts from the boxes and exposing the plants to a dry air stream for 15 minutes, which led to a fresh-weight loss of...
3.2. Modeling responses to abiotic stresses

3.2 Modeling responses to abiotic stresses

3.2.1 Data pre-processing

Chip normalization

As outlined in section 1.2, microarray experiments are subject to a number of systemic variations. To correct for those, a number of different normalization methods have been proposed and it has been shown that the employed normalization method can greatly affect the outcome of e.g. detection of differentially expressed genes [Hoffmann et al., 2002]. The Arabidopsis Information Resource (TAIR) offers download of the AtGenExpress data sets both as raw CEL files and as text files normalized using the Affymetrix Microarray Suite (MAS5.0) method. To assess which normalization method is best suited to the data set at hand, we compared observed differences between the two replicates contained in the AtGenExpress data after using the Robust Multi-array Analysis with correction for GC content of the oligo (GCRMA [Wu and Irizarry, 2007], MATLAB function affygcrma from the Bioinformatics Toolbox) or MAS5.0 normalization, respectively. We computed Spearman’s rank correlation coefficients for individual replicate array pairs taking into account all measured probe sets to evaluate the effect the normalization procedure had on replicate consistency. Table 3.1 lists the average replicate correlations attained on different AtGenExpress abiotic stress data sets. Both normalization methods achieve near-perfect replicate correlations. Consistent with the findings of [Harr and Schloetterer, 2006], in an evaluation over all replicate pairs of the abiotic stress data sets, Spearman correlations between replicate samples were, however, significantly higher after GCRMA normalization than after MAS5.0 normalization ($p = 1.5 \times 10^{-6}$, Mann-Whitney U test).

Correlations computed between replicate samples using all probe sets help evaluating the capability of a normalization method to produce replicate-consistent data in a classical two-dimensional experimental design (genes * replicated measurements of a number of experimental conditions). The AtGenExpress data set has, however, an underlying three-dimensional experimental design (genes * time points * stress conditions). We thus proceeded to analyze the effect of normalization procedures on the correlation between replicated gene expression profiles over time. Already in very early attempts at analyzing genome-wide transcriptomics data it has been noted that poorly expressed genes tend to have low signal-to-noise ratios (e.g. [Beißbarth et al., 2000]). Consequently, lowly expressed genes can not be expected to show reliable correlation between replicates. We thus focused our evaluation on the 25% most highly expressed genes (as measured by their mean expression level over both replicate time-courses) in the respective data sets. Table 3.2 shows the average Spearman correlations computed over the replicate time-course profiles attained in the respective data sets. As can be seen from these values, it is considerably harder to achieve high correlations on replicated time-courses as compared to array-wide correlations between replicate samples. The difference between GCRMA and MAS5.0 normalization, however, was more pronounced in the time-course based anal-
3.2. Modeling responses to abiotic stresses

ysis, with GCRMA consistently outperforming MAS5.0. Consequently, when compared over all abiotic stresses, GCRMA performs significantly better than MAS5.0 in terms of replicate time-course correlation ($p = 1.4 \times 10^{-52}$, Mann-Whitney U test). Figure 3.3 shows the distribution of time-course correlation coefficients for highly expressed genes in the respective abiotic stress data sets as violin plots. As can be seen from the violin plots, the majority of highly expressed genes have high correlations over the individual stress time-courses. However, in all data sets there are also outliers showing weak or even negative replicate correlations. Those outliers are most likely genes showing very stable expression with small random fluctuations, which do not show the same direction in the two replicates.

Based on the above evaluations, we decided to use GCRMA-normalized data for the remainder of our analysis.

<table>
<thead>
<tr>
<th>Data set</th>
<th># Correlations</th>
<th>Avg corr GCRMA</th>
<th>Avg corr MAS5.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heat</td>
<td>8</td>
<td>0.997</td>
<td>0.995</td>
</tr>
<tr>
<td>Drought</td>
<td>7</td>
<td>0.995</td>
<td>0.992</td>
</tr>
<tr>
<td>Cold</td>
<td>6</td>
<td>0.993</td>
<td>0.988</td>
</tr>
<tr>
<td>Osmotic</td>
<td>6</td>
<td>0.995</td>
<td>0.99</td>
</tr>
<tr>
<td>Salt</td>
<td>6</td>
<td>0.989</td>
<td>0.984</td>
</tr>
</tbody>
</table>

Table 3.1: Average Spearman correlation coefficients between replicate samples in the AtGenExpress abiotic stress root tissue data sets

<table>
<thead>
<tr>
<th>Data set</th>
<th>Time points</th>
<th># Correlations</th>
<th>Avg corr GCRMA</th>
<th>Avg corr MAS5.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heat</td>
<td>8</td>
<td>5702</td>
<td>0.792</td>
<td>0.756</td>
</tr>
<tr>
<td>Drought</td>
<td>7</td>
<td>5702</td>
<td>0.575</td>
<td>0.512</td>
</tr>
<tr>
<td>Cold</td>
<td>6</td>
<td>5702</td>
<td>0.536</td>
<td>0.495</td>
</tr>
<tr>
<td>Osmotic</td>
<td>6</td>
<td>5702</td>
<td>0.648</td>
<td>0.6</td>
</tr>
<tr>
<td>Salt</td>
<td>6</td>
<td>5702</td>
<td>0.617</td>
<td>0.591</td>
</tr>
</tbody>
</table>

Table 3.2: Average Spearman correlation coefficients between replicate time-courses of the 25% most highly expressed genes in the AtGenExpress abiotic stress root tissue data sets

Selection of genes suitable for modeling

While it can be regarded sufficient to select genes solely according to their differential expression for observational studies as e.g [Kilian et al., 2007], in-depth modeling of biological processes is highly dependent on the selection of robust and relevant data. To select genes with robust information, relevant to the processes to be modeled, we thus devised the following set of gene selection criteria:

1. Concordant replicate time-courses

We computed Pearson correlation between replicate profiles of a given gene over
3.2. Modeling responses to abiotic stresses

Figure 3.3: Replicate time-course correlation distributions over the 25% highest expressed genes after GCRMA (green) and MAS5.0 (blue) normalization. Distributions of correlation coefficients are shown as violin plots. The 25th percentile, median and 75th percentile of the distributions are indicated by red lines.

differential (stress - control) expression profiles. P-values for significant increase of the measured correlations over 0 were computed using a Student’s t distribution for a transformation of the correlation, which can be shown to yield an exact p-value under the assumption of normality of the two expression profiles to be compared. Genes for which the null hypothesis of low correlation could be rejected at a significance threshold of $\alpha = 0.01$, were deemed sufficiently replicate-correlated for further analysis.

2. Robust expression or relevant GO annotation
Genes not showing an absolute intensity value higher than 6.0 for any of the time points in either control or stressed conditions were excluded from the analysis. As some relevant regulators of stress responses can be efficient already at very low concentrations, we excepted genes annotated as either transcription factors (GO terms GO:0032583 and GO:0003700), RNA-binding molecules (GO:0003723) or protein kinases (GO:0016301) from the above rule. Genes annotated with a descendant of the above mentioned terms according to the Gene Ontology (GO) [Ashburner et al., 2000] tree were considered potential regulators. The Ontology structure was downloaded from the GO website on the 14th of April 2013 and GO annotations were retrieved from the annotation file curated by The Arabidopsis Information Resource (TAIR, released 7th of March 2013).

3. Differential expression
In order for a gene to be relevant for modeling a given stress, it has to show a clear reaction to the applied stress. We thus required selected genes to have a
deregulation amplitude of at least one log2 unit (at least a two-fold change versus control). Deregulation amplitude was measured as the maximum change of stressed versus control condition in terms of mean expression over replicates.

Only genes fulfilling all of the three above rules in a given stress data set were selected as an input to modeling. Modeling itself was, unless noted otherwise, carried out on differential expression kinetics against control averaged over the two replicate samples.

Kilian et al., 2007 selected three stresses with very different characteristics for their analysis with the goal to unravel a set of core stress response genes, commonly reacting to a number of different stresses. Here, however, we selected two classes of stress treatments in order to subject them to two different mathematical modeling approaches for discovery of transcriptional response networks. The first class of data sets consisted of stresses applied in a persistent manner over the whole 24 hour measurement period, while the second class was composed of transiently applied stresses.

### 3.2.2 Persistently applied stresses (cold, osmotic and salt stress)

For the analysis of persistently applied stresses, we focused on the cold, osmotic and salt stress data sets. Those three stresses were selected because the mode of stress treatment between them was similar, as all stresses were applied by putting plants into modified (cooled down, addition of mannitol or NaCl, respectively) medium. The cold stress, however, also contained direct stress exposure of the leaves, as the plants’ boxes were first put on ice (which should rapidly cool down the growth medium) and then transferred to a cold room (4°C). Thus, owing to the mode of stress application, osmotic and salt stress (and possibly also cold stress) were expected to be sensed first by the root tissue. Furthermore, salt and osmotic stress, as well as cold stress to a lower degree, lead to changes in osmolarity of the medium which can be detected by the plants’ roots. Consistent with this, previous work using cDNA chips measuring ~8,000 genes, identified around 30% of all observed genes as having a common expression in response to cold, osmotic and salt stress treatments (Kreps et al., 2002; Seki et al., 2002), suggesting that — in addition to stress-specific responses — a common set of overlapping stress response components is triggered.

Application of the gene selection criteria detailed in the previous section to the cold, osmotic and salt stress data sets led to selection of 4,174 genes retained in at least one of the root tissue data sets and of 5,707 genes retained in the corresponding shoot tissue data sets. Figure 3.4 shows Venn-diagrams of the selections in root and shoot tissue data sets. While the stresses were applied primarily to the root tissue, the number of genes reacting to the stress in the shoot tissue was higher for both cold and osmotic stress, while the salt stress lead to less deregulation in the shoot than in the root tissue. As expected, due to the osmolarity change induced by adding salt to the medium, the overlap between salt and osmotic stress-regulated genes was large (~40% of salt-regulated genes were contained in the response of root tissue to osmotic stress). In the shoot tissue, more than 80% of salt-regulated genes were also deregulated by osmotic stress, indicating that the systemic signal sent from the root to the shoot tissue upon salt stress perception might be dominated by an osmolarity-change component. Generally, the overlap of genes responding to
the different stresses was larger in the shoot tissue than in the root tissue (e.g. only ~20% of overlap between cold and osmotic stress in the root tissue, while the overlap in the shoot tissue was above 40%). This can be interpreted as resulting from relatively specific stress responses being initiated in the primary tissue of stress exposure (root) which then lead to less specific systemic signaling to the shoot tissue, and thus to initiation of more overlapping general stress responses.

Looking at the timing of the stress responses in the different tissues revealed that, under osmotic and salt stress, more than 20% of selected genes showed early (within 30 minutes to one hour) responses to the applied stress in the root tissue. The response to the same stresses in the shoot tissue was much slower (c.f. Table 3.3) with the main onset of stress responses being measured after three to six hours. In contrast to this, the cold stress elicited slightly more early responses in the shoot than in the root tissue and was, in both tissues, characterized by a gradually increasing number of genes responding to the stress over the measured time-course. This delayed stress response is likely due to the rates of molecular reactions being slowed down by the low temperature [Kilian et al., 2007].

Taken together, these results indicate that salt and osmotic stress are first sensed in the root tissue, which induces an early stress response and initiates systemic signaling leading to responses in the shoot tissue being induced only three to six hours after stress application. The cold stress is sensed by both tissues, leads to relatively few early gene responses, and is characterized by a slowly increasing number of responding genes over the whole 24 hour measurement period.

Figure 3.4: Overlap of genes selected for modeling in the root tissue (left panel) and shoot tissue (right panel) data sets for cold (red), osmotic (green) and salt (blue) stress according to the criteria detailed in section 3.2.1 on page 47.
3.2. Modeling responses to abiotic stresses

<table>
<thead>
<tr>
<th>Condition</th>
<th>30min</th>
<th>1h</th>
<th>3h</th>
<th>6h</th>
<th>12h</th>
<th>24h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cold root</td>
<td>1.3</td>
<td>2.1</td>
<td>9.2</td>
<td>17.9</td>
<td>31.6</td>
<td>38.0</td>
</tr>
<tr>
<td>Cold shoot</td>
<td>5.3</td>
<td>5.6</td>
<td>11.8</td>
<td>15.0</td>
<td>30.9</td>
<td>31.4</td>
</tr>
<tr>
<td>Osmotic root</td>
<td>7.7</td>
<td>13.5</td>
<td>16.3</td>
<td>21.8</td>
<td>31.1</td>
<td>9.6</td>
</tr>
<tr>
<td>Osmotic shoot</td>
<td>1.1</td>
<td>4.4</td>
<td>23.8</td>
<td>26.2</td>
<td>24.9</td>
<td>19.5</td>
</tr>
<tr>
<td>Salt root</td>
<td>6.0</td>
<td>16.9</td>
<td>26.0</td>
<td>31.9</td>
<td>13.4</td>
<td>5.8</td>
</tr>
<tr>
<td>Salt shoot</td>
<td>1.9</td>
<td>1.8</td>
<td>35.7</td>
<td>16.0</td>
<td>9.9</td>
<td>34.7</td>
</tr>
</tbody>
</table>

Table 3.3: Response onset times under cold, osmotic and salt stresses in the root and shoot tissue data. Shown are the percentage of genes with the earliest deregulation (FC \( \geq 2 \) or \( \leq 0.5 \)) occurring at each given time point.

Shiraishi model

[Shiraishi et al., 2010] developed a so-called state-space model to learn an underlying regulatory network from measured expression data under different stresses. Figure 3.5 shows a graphical scheme illustrating the basic idea behind their model: Genes are grouped into clusters that follow a similar expression profile and a set of regulatory interactions best explaining the observed cluster intermediate profiles is learned from the data. Pooling of genes into clusters and estimation of regulatory interactions between clusters instead of genes largely reduces the complexity of the computational problem and the influence of random fluctuations in individual genes.

Model formalization according to [Shiraishi et al., 2010]

The state-space model is characterized by two mathematical models. The observational model describes the dependency of observed expression of genes \( y \) under stress \( s \) and at time point \( t \) from the expression profiles \( x \) of their associated cluster intermediates by using the cluster assignment matrix \( C \). \( w \) indicates experimental noise, distributed according to \( N(0, \sigma^2 I) \), which is restricted to a global estimate of stochastic noise given by \( \sigma^2 \) and assumes independence of noise between genes (no non-zero co-variance terms):

\[
y_{s,t} = C * x_{s,t} + w_{s,t} \tag{3.1}
\]

The system model describes rules for computing the expression of cluster intermediates \( x \) at time \( t \) from the preceding time point \( t-1 \) by applying regulatory interaction rules between cluster intermediates (encoded by the matrix \( A \)) as well as the direct influence (vector \( b_s \)) applied stresses have on the individual cluster intermediates. \( v \) encodes the system-level noise and is distributed according to \( N(0, \tau^2 I) \), again restricted to a global estimation of system noise by the parameter \( \tau^2 \):

\[
x_{s,t} = A * x_{s,t-1} + b_s + v_{s,t} \tag{3.2}
\]

Notably, this model formalization makes a couple of assumptions about the nature of the modeled system:

1. Stress influences are constant over the full time course
3.2. Modeling responses to abiotic stresses

Figure 3.5: Overview of the state space model defined by [Shiraishi et al., 2010]: Genes are grouped into clusters according to their expression profiles and then represented by cluster intermediate profiles. The regulatory network learned is composed of regulatory interactions between individual cluster intermediates as well as direct influences of external stresses on cluster intermediates. This figure is a reproduction of figure 1 from [Shiraishi et al., 2010].

2. The underlying regulatory network wiring is identical in response to the different stresses. The expression level of cluster intermediates, however, determine which regulation rules influence the system most, thus allowing for differential behavior of the system under different stresses.

3. Cluster intermediate expression at a given time point is only dependent on the immediate previous time point and not influenced by other past time points.

4. The degree of experimental noise is identical for all individual genes and all individual cluster intermediates and there is no noise co-variance between genes or cluster intermediates.

5. The regulatory network is characterized by relatively few interactions, leading to a sparse regulation matrix $A$.

While these assumptions may seem very restrictive, they are needed in the step of model fitting, as relaxing any of them would significantly increase the number of parameters to
3.2. Modeling responses to abiotic stresses

be estimated from the data. It can be argued that for any system-scale biological data set, which tend to measure only short time courses with limited numbers of replicates, a relaxation of the above restrictions would lead to an under-determined system.

Model fitting devised by Shiraishi et al. The estimation of model parameters from data, commonly referred to as model fitting, depends on the definition of a score measuring to which degree a model, with a given set of parameter values \( \theta \), is congruent with observed data. Model fitting algorithms then search for parameter sets \( \theta \) yielding good scores. Shiraishi et al. use an Expectation-Maximization (EM) algorithm for model fitting. EM algorithms consist of two steps: In the E-step, the model fitness, given a set of model parameters (initialized randomly at first) is computed. In the M-step, the individual parameters are changed, one at a time, in order to improve the overall model fitness. The two steps are then repeatedly executed until no significant improvement of the model fitness can be achieved anymore.

More specifically, the EM algorithm laid out in the supplementary material of [Shiraishi et al., 2010] starts with random initialization of \( \theta \) (including, notably, a random assignment of genes to clusters) and consists of the following steps:

- **E-step** As the expression profiles of cluster intermediates are not directly observed in the data, they have to be estimated alongside the remaining model parameters \( \theta = (A, C, \{b_s\}, \sigma^2, \tau^2) \). In the MATLAB implementation of their model provided by Shiraishi et al., \( x_{s,t} \) is estimated using a Kalman filter and smoother. Kalman filtering is a method developed in the field of control theory. Given a set of regulatory rules and gene expression measurements at some of the modeled time points, it computes the optimal expression trajectory in terms of agreement with both the regulatory rules and measured expression.

The estimated cluster intermediate profiles are used to compute a fitness function. They use a log-likelihood function \( L \), representing the log-transformed theoretical probability of the data being reproduced by the model in its current parameterization. \( L \) is computed as a sum of three penalty terms. The first term penalizes quadratic deviation of individual genes from their associated cluster intermediates:

\[
- \frac{1}{2\sigma^2} \sum_{s=1}^{S} \sum_{t=1}^{T} (y_{s,t} - Cx_{s,t})' (y_{s,t} - Cx_{s,t})
\]  

The second term penalizes quadratic deviation of the cluster intermediate profiles learned by the Kalman filter from expression values computed according to simply applying the rules of the regulatory network:

\[
- \frac{1}{2} \sum_{s=1}^{S} \sum_{t=1}^{T} (x_{s,t} - Ax_{s,t-1} - b_s)' (x_{s,t} - Ax_{s,t-1} - b_s)
\]  

The last term penalizes large noise estimates:

\[
- \frac{NST}{2} \log(\sigma^2) - \frac{KS(T-1)}{2} \log(\tau^2)
\]  

53
3.2. Modeling responses to abiotic stresses

- **M-step** The M-step first optimizes cluster assignment of genes according to the newly learned cluster intermediate profiles. After gene reassignment, regulatory rules are updated and the noise parameters re-computed.

The model fitting code furthermore enforces a sparsity constraint on matrices \( A \) and \( b \) to assure that "simple" regulatory network solutions (i.e. less regulatory interactions) are preferred to more complex networks. To avoid being trapped in local optima, Shiraishi et al. have devised a cluster split-and-merge procedure. After an initial EM-run has converged, they select triplets of clusters, where one cluster shows low fitness in terms of the agreement between gene profiles and the cluster intermediate profiles and the two other clusters show similar expression profiles. They then — using a greedy heuristic — select the four most promising split-and-merge candidates. For each of those four candidates, a split-and-merge is carried out and the resulting model is optimized using the EM-algorithm outlined above. If one of the four candidates produces a better result than the solution found after the first EM-run, the split-and-merge procedure is repeated, otherwise the solution resulting from previous EM-run is returned.

Shiraishi results

**Determination of number of clusters** An important input to the Shiraishi method is the number of clusters, \( K \), the genes should be classified into. In their paper, Shiraishi et al. propose trying a range of different values for \( K \) and using a score introduced in [Krzanowski and Lai, 1985] to evaluate which number of clusters is optimal. The score is defined as follows:

\[
KL(K) = \frac{\text{Diff}(K)}{\text{Diff}(K+1)}
\]  

(3.6)

with

\[
\text{Diff}(K) = (K - 1)^{2/T \times N} \sigma_K^2 - K^{2/T \times N} \sigma_K^2
\]  

(3.7)

Since fitting the model to the stress data studied in this project took between two days and two weeks — depending on the complexity of the system — on a high-performance multicore machine, running it for multiple values of \( K \) was not feasible. We thus determined the optimal number of clusters separately from the Shiraishi model. To this end, we ran k-means clustering using euclidean distance for a range of \( K \) between 2 and 50 and replaced the estimate of observation noise \( \sigma^2 \) by the sum of squared deviations of individual genes from their associated cluster centroids. Choosing \( \text{argmax}_K KL(K) \) led to 35 clusters for the root and 45 for the shoot data (plots of the KL score for \( K \) ranging from 3 to 49 are shown in supplementary figures A.3 and A.4).

**Evaluation of parameter initialization** In their paper Shiraishi et al. propose to start the model fitting with a random initialization of cluster assignments. They argue that the learning of regulatory rules and the assignment of clusters are interdependent and should be tackled in unison. This might be a good strategy for modeling small, simple systems as they have shown in their paper. Applying the Shiraishi method to our data resulted, however, in rather poor clustering of genes and a low final likelihood score \( L_{\text{root.rand}} = -107,287 \). We thus adapted their code to allow supplying an initial clustering. When supplying a cluster assignment matrix \( C \) determined by running kmeans
3.2. Modeling responses to abiotic stresses

clustering using euclidean distances on a concatenation of the expression kinetics measured under the three stresses, the method converged to its final solution much faster (2.6 days versus around 10 days with random initialization for the root tissue data sets) and resulted in a higher likelihood score ($L_{root,kmeans} = -59,383$). The same tendencies were observed on the shoot data set ($L_{shoot,rand} = -141,428$; $L_{shoot,kmeans} = -84,023$). While the initial clustering was determined using an external clustering procedure, the Shiraishi model can adapt gene assignments during its fitting procedure. For both the root and shoot data, the optimization procedure resulted in only small changes in the clustering as indicated by an F-measure of 0.97 and 0.94 respectively, obtained when comparing the initial clustering to the clustering returned after parameter optimization.

The above results indicate that initializing the method with a predetermined clustering improves the final solution both in terms of the clustering and in terms of the consistency of learned regulatory interactions while still allowing the optimization procedure to adapt cluster assignments where it is beneficial to the global solution of the problem.

Predicted regulatory network: root tissue

The regulatory network structure inferred by the Shiraishi model initialized with kmeans clustering on the root tissue data is shown in Figure 3.6. The direct influences salt and osmotic stress are inferred to exert on

![Figure 3.6: Inferred regulatory network for the root tissue data. Clusters are represented as nodes, with node size indicating the number of contained genes. Strong regulatory interactions (absolute coefficients of at least 0.2) are depicted by edges (red: down-regulation; green up-regulation). Edge width indicates the strength of inferred regulation.](image)

the individual clusters are very similar both in terms of which clusters are influenced and in magnitude of those influences (see Figure 3.7). Furthermore, salt stress additionally to sharing all the direct influences of the osmotic stress, induces a number of additional clusters. Cold stress is predicted to have less direct influences on the cluster intermediate
3.2. Modeling responses to abiotic stresses

Figure 3.7: Fitted regulation matrices for the root tissue data. Shown are the regulation matrix \( A \) (left) and the direct stress influence matrix \( b \) (right). The individual coefficients are represented by color-coded cells. Blue (red) cells indicate negative (positive) coefficients.

expressions. Generally, the expression kinetics of cluster intermediates are characterized by early deregulations caused by the direct stress influences and a relaxation to basal or alternative, stable levels of expression at the end of the measured time-course. The system is in a steady-state at the end of the measured expression as extrapolation of the expression kinetics according to the inferred regulation rules showed little to no changes from the state attained at the 24 hour measurement.

Two rather small clusters are induced most strongly by the applied stresses: cluster 20 (42 genes) by cold and osmotic stress and cluster 35 (11 genes) by all three of the stresses. Furthermore, they are predicted to be part of negative (e.g. Cluster 20 → Cluster 35 → Cluster 11 → Cluster 20) and positive (Cluster 35 → Cluster 11 → Cluster 23 → Cluster 35) feedback loops, which have been shown to be common features in regulatory response networks (e.g. negative feedback in NFκB signaling [Krishna et al., 2006]; review on positive feedback circuits [Mitrophanov and Groisman, 2008]). The only other direct influence of cold stress is predicted to be a repression of the 8 genes assigned to cluster 22. The clusters strongest influenced by salt but not osmotic stress, apart from cluster
3.2. Modeling responses to abiotic stresses

23, are cluster 31 (73 genes) and cluster 30 (95 genes).

The moderate estimated gene-level noise with variance $\sigma^2$ of 0.22 indicates that cluster intermediate expression kinetics are globally consistent with the kinetics of genes assigned to the respective cluster intermediates. Inspection of the clusters which take part in the central feedback loops (clusters 20, 35, 11 and 23) and thus significantly shape the stress response revealed, however, significant deviations of measured gene expression from the predicted cluster intermediates (Figure 3.8). Furthermore, the cluster intermediates learned from the data using the Kalman filter procedure (blue lines in Figure 3.8), show strong oscillations and deviate substantially (especially for clusters 23 and 35) from cluster intermediate profiles predicted by applying the inferred regulatory rules (green lines). The predicted profiles show less but still significant oscillations at the cost of not matching the gene expression data as closely. Cluster 35 shows early and high induction in all three stresses, while the induction of cluster 20 takes longer under cold than under salt and osmotic stress, congruent with the strengths of predicted stress influences. The behavior of cluster 35 is mimicked, with a smaller amplitude, by cluster 11. Cluster 23, in turn, shows similar expression kinetics to cluster 11 under cold and osmotic stress, but is more strongly induced under salt stress, which in the model leads to inference of a direct positive regulation by salt stress.

The clusters directly influenced by cold (cluster 22), directly influenced by salt but not osmotic stress (clusters 31, 30) as well as another cluster not connected to the central feedback loops (cluster 3, 29 genes) show less oscillations in their learned and predicted cluster intermediates (Figure 3.9) as well as a better agreement between measured gene

---

Figure 3.8: Expression of four clusters taking part in feedback loops. For more detailed descriptions see accompanying text.
3.2. Modeling responses to abiotic stresses

expression data and cluster intermediates. Cluster 22, consistent with its predicted di-
rect repression by cold stress is down-regulated early under cold and its initial expression restored after 12 hours through the influence of cluster 3. Under salt and osmotic stress, it is up-regulated first at the 6 hour time point, again mediated by the gradual down-
regulation of cluster 3. Cluster 31 and 30 show little response to cold and osmotic stress, but are strongly upregulated under salt stress with similar expression trajectories.

The differences in terms of agreement with measured gene expression data and cluster intermediate oscillations apparent between clusters shown in Figures 3.8 and 3.9, respectively, are representative for the remaining clusters, with clusters part of or downstream of feedback loops showing more oscillations and less agreement with gene expression data than clusters either regulated solely by stress influences or on simple, linear regulatory pathways.

In order to be able to assess biological functions of individual clusters and judge the biological plausibility of regulatory rules inferred by the model, we performed GO-enrichment analysis of the identified clusters.

**Functional analysis** The same releases of the GO tree and the GO annotation files as listed in section 3.2.1 were used for GO enrichment analysis. We focused the GO analysis on the biological process aspect of the GO. To this end, we completed each gene’s annotation by adding all eventually missing ancestor terms of already annotated terms. We then computed p-values for enrichment of GO-term annotations in individual clusters by using a hypergeometric distribution with the background set composed of all genes selected for modeling. The p-values were corrected for multiple testing using the
3.2. Modeling responses to abiotic stresses

Benjamini-Hochberg method and GO terms significant at a level of $\alpha \leq 0.01$ were deemed significant. In the following, we will list and interpret GO annotations enriched in the main clusters of interest as well as individual genes known to be involved in stress responses and their regulation:

**Central regulatory clusters**

Cluster 35 contains the three main cold transcription factors CBF1, CBF2 and CBF3 as well as one of their known targets RD29A. Consequently, it is enriched for cold response genes. Additionally, a galactinol synthase GolS2, involved in the synthesis of raffinoses which have anti-oxidative effects [Nishizawa et al., 2008], and the heat shock factor HSFA6B, which is known to be induced under salt and drought stresses are contained in cluster 35.

The second cluster indicated as a main mediator of salt and osmotic stress responses and predicted to repress cluster 35, cluster 20, is enriched for a number of general stress response terms, like e.g. response to water deprivation (11/42 genes), hyperosmotic response (8/42) and response to ABA (11/42). It furthermore contains 15 genes annotated as transcription factors. Contained genes include a $\beta$-glucosidase BGLU24, involved in the re-activation of ABA from its conjugated form, two MAP3Ks, MAPKKK17 and 18, which have been shown to be implicated in ABA-mediated long-term responses to dehydration stress and are activated by transcriptional induction upon ABA, drought and salt stress [Danquah et al., 2013a]. DREB2A and NAC019, which are canonical examples of the two ABA-independent drought/osmotic stress pathways, are also contained in cluster 20. Furthermore, RAP2.6, a transcription factor downstream of the CBFs is contained in cluster 20, which does not fit with the predicted regulation of cluster 35 by cluster 20. There are no genes known to be implicated in the negative regulation of CBFs contained in cluster 20.

Cluster 11, which is predicted to be up-regulated by cluster 35, is enriched for transcription factors (22/82) as well as genes responsive to ABA, salt and osmotic stress (11/82). The promoters of its genes are enriched for ABRE motifs (38/82, $p = 10^{-8}$ according to the Athena tool [O’Connor et al., 2005]), but not for the CRT/DRE motif which would have been expected for targets of CBFs. Prominent members of cluster 11 are the ethylene response factors ERF4 and ERF8 as well as two heat shock factors HSFA1E and HSFB2A.

Cluster 23, which is downstream of cluster 11 and additionally predicted to be directly regulated by salt stress, is also enriched for transcription factors (9/28) and genes responding to the plant hormone jasmonic acid (6/28) including three jasmonate zim domain proteins JAZ5, JAZ7 and JAZ10 which have been shown to be transcriptional repressors of a number of targets including the transcription factors MYC2, 3 and 4 thus mediating jasmonic acid signals [Kazan and Manners, 2012]. The predicted regulations exerted by Cluster 23 are, however, all positive regulations and none of the known JAZ targets are contained in the clusters predicted to be regulated by it.

**Stress-specific clusters**

Cluster 3 is down-regulated under salt and osmotic stress. It is enriched for genes involved in glycosinolate, a plant pesticide, biosynthesis (5/29). No canonical abiotic stress response or regulatory genes are contained in the cluster.
3.2. Modeling responses to abiotic stresses

The predicted repression target of cluster 3, cluster 22, is composed of 8 heat shock chaperones and consequently enriched for genes responding to heat stress (8/8). There is no known literature linking genes of cluster 3 to downregulation of those HSPs.

Cluster 31 is predicted to be directly induced by salt stress only. It has no significantly enriched GO annotations.

Cluster 30, which is predicted to be activated by cluster 31, is enriched for signal transduction genes (28/95) from the jasmonic acid (13/95) and salicylic acid (13/95) signaling pathways and furthermore contains three annexins, which are calcium-dependent phospholipid binding proteins implicated in calcium signaling upon oxidative stress [Richards et al., 2014].

As shown in the previous paragraphs the inferred regulatory rules do conform with the predicted cluster intermediate expression profiles. Functional enrichment analysis and literature review for individual key regulators of abiotic stress responses did, however, show that in terms of annotated biological functions, most clusters do not show a single clear function. Furthermore, there was no indication of known key regulatory processes (like e.g. the CBF-dependent regulation of cold response genes) being correctly inferred by the model.

**Predicted regulatory network: shoot tissue** Given the later onset of salt and osmotic stress responses in the shoot tissue (c.f. Table 3.3), we expected the model to have more problems in predicting the direct stress influences on the shoot tissue than on the root tissue. Consistent with this assumption, the model predicted far fewer and weaker direct stress influences than in the root tissue, especially for the salt and osmotic stress. The predicted direct cold stress influences were similar in number and magnitude to the root situation (c.f. Figure 3.10). Lacking strong direct stress influences, the observed gene responses were instead modeled to be driven by oscillation-inducing network motifs consisting of two clusters reciprocally activating and repressing each other strongly (e.g. clusters 40 and 32, 14 and 10 shown in Figure A.5 in the appendix). Consequently, the learned and predicted cluster intermediate time courses showed a high degree of oscillatory behavior and did not attain a second steady-state at the end of the measured time course.

This highly oscillatory expression behavior is not plausible biologically and we decided to not further investigate the biological functions assigned to individual clusters or biological plausibility of the inferred regulations.

**Discussion of modeling results**

The Shiraishi state-space model has been shown to be able to infer valid regulatory networks on data measured on medium-scale (<1000 genes) and medium-complexity (<10 clusters) biological systems measured at eight to twelve time points [Shiraishi et al., 2010]. When fit to appropriate data, it allows prediction of the effect a combined application of external stresses has on the modeled system. The model is designed to reduce the complexity of the system by pooling expression data of individual genes. One of the
3.2. Modeling responses to abiotic stresses

Figure 3.10: Fitted regulation matrices for the shoot tissue data. Shown are the regulation matrix $A$ (left) and the direct stress influence matrix $b$ (right). The individual coefficients are represented by color-coded cells. Blue (red) cells indicate negative (positive) coefficients.

A key assumption of the model is that the same regulatory circuit underlies all measured responses. While the model fit to the abiotic stress data set, did yield general features consistent with the expected behavior of the system (more direct stress influences modeled for stresses leading to strong and early gene responses), it did not produce biologically plausible regulatory interactions between clusters. This could be due to a number of factors:

The assumption of equal regulatory kinetics in all measured responses might be violated by the slowing of enzymatic reactions at low temperatures. Fitting the model only to osmotic and salt stress responses did, however, not lead to biologically more plausible regulatory interactions. The selected genes, while carefully tested for robust and relevant behavior, could contain too many genes not relevant to the establishment of regulatory interactions. Reduction of the selected genes to only include transcription factors, did, however, not significantly change the conclusions drawn above. The fitting of the model could have suffered from too few measured data points, leading to an under-determined system, for which many model parameterizations would lead to good likelihood scores.
3.2. Modeling responses to abiotic stresses

This should, however, not be the case as the ratio of measured data points to free parameters is about 14:1 even without considering that the enforced sparsity constraint should eliminate additional free parameters. It has to be noted, however, that the total number of measured data points does not necessarily reflect the amount of information contained in the data, which depends strongly on the number of contained distinguishable expression profiles as well as on the consistency between expression profiles of individual genes. Furthermore, few-sample, high-complexity data as the one used here, can lead to overfitting. Finally, the analyzed data set, while representing the largest existing, experimentally consistent, data collection measuring time-resolved responses to abiotic stresses of *Arabidopsis*, contains only seven time points. Although the time points are irregularly spaced and focused on early reactions to the stresses, it is possible that the data does not allow distinction of the individual responses in terms of onset time of regulators and their immediate targets. This assumption is backed by the fact that the model places the main cold regulation factors CBF and one of their canonical downstream targets (RD29A) in a single cluster.

Independent of those considerations, analysis of the inferred state-space model revealed a further problem: The complexity of inferred regulations as well as the reduction of the data to cluster intermediates make it difficult to interpret the results unless very clear biological processes can be attributed to each individual cluster. To date, systems-level computational modeling is almost exclusively used as a tool for hypothesis generation. The generated hypotheses are then subjected to experimental verification and their results used to adapt the initial model accordingly. In this setting, the Shiraishi model seems not to be the right choice as it does not easily facilitate generation of testable hypotheses nor inclusion of prior knowledge into the model fitting procedure.
3.2.3 Transiently applied stresses (heat and drought stress)

As laid out in section 3.1.2 on page 43, heat and drought stress, especially occurring in combination, are the abiotic stresses most impacting crop loss world-wide. Drought stress, in the field, is a gradual process as the lack of rain leads to a gradually decreasing water content in the soil. It is thus hard to define when exactly plants start to perceive the lowered water availability as a stress. In order to be able to measure the early response of Arabidopsis thaliana to drought stress, the experimental protocol used to create the AtGenExpress drought data set is set up as follows: plants are removed from their growth boxes and a dry air stream is applied for 15 minutes. Thus, a clear starting point of the drought stress can be established, and the reaction to and recovery from stress can be measured accurately. The AtGenExpress heat data set was established by putting the plants in a growth chamber heated up to 38°C for three hours, measuring gene expression during and after the application of the stress. To establish early responses to relieving the heat stress, the time-series was extended by a 4 hour (1 hour post stress) time point. The selection criteria defined in section 3.2.1 on page 47 led to the selection of 3,571 genes in the heat and 1,186 in the drought stress data set for the root tissue and to 3,566 and 1,043 genes for the shoot tissue data, respectively. Figure 3.11 shows overlaps between those sets in the root (left) and shoot (right) tissue data.

Figure 3.11: Overlap of genes selected for modeling in the root tissue (left panel) and shoot tissue (right panel) data sets for heat (red) and drought (green) stress according to the criteria detailed in section 3.2.1 on page 47.

Due to the mode of stress application, many of the genes in the heat and drought data sets show profiles of transient induction or down-regulation during stress application, followed by a fall-back to initial levels after the stress has been removed. On a global level this can be observed by the high percentage of genes showing their earliest deregulation during or just after stress exposure (c.f. Table 3.4). Around 80% of the selected genes show their
3.2. Modeling responses to abiotic stresses

<table>
<thead>
<tr>
<th>Condition</th>
<th>15min</th>
<th>30min</th>
<th>1h</th>
<th>3h</th>
<th>4h</th>
<th>6h</th>
<th>12h</th>
<th>24h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heat root</td>
<td>4.6</td>
<td>3.4</td>
<td>20.8</td>
<td>50.2</td>
<td>14.4</td>
<td>0.6</td>
<td>5.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Heat shoot</td>
<td>10.2</td>
<td>7.1</td>
<td>22.5</td>
<td>35.0</td>
<td>21.8</td>
<td>2.6</td>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>Drought root</td>
<td>11.5</td>
<td>36.0</td>
<td>34.0</td>
<td>3.0</td>
<td>ND</td>
<td>13.9</td>
<td>1.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Drought shoot</td>
<td>15.3</td>
<td>29.0</td>
<td>17.8</td>
<td>21.5</td>
<td>ND</td>
<td>12.5</td>
<td>2.0</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Table 3.4: Response onset times under heat and drought stress in the root and shoot tissue data. Shown are the percentage of genes with the earliest deregulation (FC ≥ 2 or ≤ 0.5) occurring at each given time point.

first deregulation within the first three hours of the experiment for both heat and drought stresses and in both root and shoot tissue.

Is the Shiraishi model applicable?

The Shiraishi model, presented in the previous section, assumes constant stress effects over the whole measured time-course. To evaluate whether it would be worthwhile adapting the model to incorporate non-permanent stresses, we created two artificial data sets. Two cluster intermediates influenced by two stresses were created according to the system model detailed in equation [3.2] with the following parameters:

\[
A = \begin{bmatrix} 1 & 0 \\ 0.5 & 0.8 \end{bmatrix} 
\]  \hspace{1cm} (3.8)

\[
b = \begin{bmatrix} 0.5 & -0.5 \\ 0 & 0 \end{bmatrix} 
\]  \hspace{1cm} (3.9)

We then drew 40 gene expression profiles (at time points 0, 0.25, 0.5, 1, 3, 6, 12 and 24 hours) with equal probability from the two cluster intermediates by adding Gaussian white noise with variance 0.2. We repeated training the Shiraishi model on randomly drawn genes 10 times and evaluated its capability of estimating the true regulatory network on which the cluster intermediate expressions were based. Unsurprisingly, on this very easy example, the true regulatory network was retrieved almost perfectly in each of the runs, with average deviation from the true values in \( A \) at 0.0013 ± 0.0002 and in \( b \) at 0.007 ± 0.0008.

For the second test data set, we modified the system model by including a term restricting the effect of the two stresses to the first 3 hours and 30 minutes, respectively:

\[
x_{s,t} = Ax_{s,t-1} + sp_{s,t}b_s + v_{s,t} 
\]  \hspace{1cm} (3.10)

\( sp_{s,t} \) is set to 1, if stress \( s \) is still active at time point \( t \) and to 0 otherwise. After adapting the code of the EM algorithm accordingly for model fitting on non-permanent stresses, we followed the same evaluation procedure as for the permanent stress test data set outlined above. On the transient stress toy data set, the method failed, however, to reproduce the correct regulatory network (average deviation from the true values in \( A \) at 0.1 ± 0.02 and in \( b \) at 0.82 ± 0.08). The very large error in estimation of the stress influence matrix \( b \) can
be explained by the reduced number of data points carrying information about the stress influences. Given these poor results on a very simple test case with low measurement noise, we decided not to use the Shiraishi model on the complex and much noisier system represented by the AtGenExpress heat and drought stress data sets.

The impulse model

As stated above, many genes in the heat and drought data sets show profiles of transient induction during stress application which is then followed by a transition to a steady-state after the stress has been removed. Chechik and Koller, 2009 developed a parametric model that captures such temporal responses. This impulse model explicitly represents natural temporal properties such as the onset and the offset time of the response and in their paper has been successfully used for missing value imputation in data sets measuring the response of yeast to various environmental changes. They furthermore show evidence that the binding affinity of transcription factors is in a strong relation to the activation timing of its targets as measured by the impulse model.

Model formalization according to Chechik and Koller  The impulse model captures two-transition behaviors composed of a response onset and offset and has six free parameters. Three of them describe amplitudes, namely the amplitude before stimulation ($h_0$), the amplitude at the peak ($h_1$) and the amplitude at the final steady state ($h_2$). Two parameters describe the timing of the onset ($t_1$) and offset ($t_2$) of the impulse, as the time where the rise or fall of the curve is maximal. The final parameter, $\beta$ describes the slope of both the impulse onset and offset curve. A visual representation of the model and its parameters is shown in Figure 3.12. Mathematically, the impulse model function $f$ is formalized as the product of two sigmoid functions:

$$f_\theta(x) = \frac{1}{h_1} s(x, t_1, h_0, \beta) * s(x, t_2, h_2, -\beta)$$ (3.11)

with the scaled sigmoid function $s$ defined as:

$$s(x, t, h, \beta) = h + (h_1 - h)S(\beta, t)$$ (3.12)

and the unscaled sigmoid function $S$ defined as:

$$S(\beta, t) = \frac{1}{1 + e^{-\beta(x-t)}}$$ (3.13)

Gene-pair fitting

In their paper, Chechik and Koller fit the impulse model to the expression profiles of individual genes for missing value imputation. As our goal was to construct regulatory gene networks responding to heat and drought stresses, we devised a somewhat different fitting approach. Our approach consists of identifying individual gene-pairs that are likely candidates for showing regulatory interactions and then combining them into a response network. A gene-pair composed of a regulator and its target should show a similar but time-shifted response to the stress, with the regulator reacting first and the target gene reacting with a similar shape of response but a delay in the onset of the response.
3.2. Modeling responses to abiotic stresses

Influence of fitting robustness  In order to find such regulator-target pairs, one possible strategy would be to fit two separate impulse models to the regulator and target and then compare their shapes post-fitting (separate fitting). The main problem with such a strategy is that — given the limited number of measured time points and their unequal spacing — there often exist multiple fitting solutions with very similar scores (see Figure 3.13 for an example). Thus, when choosing the optimal fit independently for the regulator and the target gene, one might end up with a low score when comparing the two fitted models’ shapes. A solution to this problem is to fit a single impulse model (concurrent fitting) to both the regulator and the target gene expression, thus making sure that poor fitting robustness does not interfere with a score measuring the similarity between the responses of the two genes (for a graphical illustration of this idea, see Figure 3.14). Consequently, we adopted a concurrent fitting procedure that finds a single impulse model fitting both the regulator and target gene expression kinetics well. An additional benefit of this approach is that the obtained fitting score can be interpreted as a direct measurement of how similar the expression kinetics of regulator and target gene are.
3.2. Modeling responses to abiotic stresses

Figure 3.13: Impulse model fitting robustness: Depending on the measured expression kinetics, multiple impulse model fitting solutions can show very similar scores. The expression of two genes (black crosses) are shown along with the optimal (thick blue line) and the next best 20 impulse model fits obtained by local optimization using a gradient descent approach starting from 200 random initializations (thin colored lines). Panel A shows a gene with a unique solution (the top 20 fits are virtually identical). Panel B shows an expression kinetic for which multiple fitting solutions fit the data equally well but show substantial variability in the response shape.

Figure 3.14: Concurrent versus separate impulse model fitting: To scan for potential regulator-target gene pairs well described by impulse models, it is beneficial to fit both genes with a single impulse model (bottom) instead of fitting them separately and trying to find similarities between time-shifted versions of the two models (top). Starting from a random model initialization for regulator and target gene (blue crosses), separate model fitting can lead to substantially different impulse shapes (orange circles and red and blue impulse shapes at the top of the figure). By fitting both genes with a single impulse model (green circle and red and blue impulse shapes at the bottom of the figure), such effects are precluded.
3.2. Modeling responses to abiotic stresses

Gene-pair fitting procedure In the following, we will describe the individual steps our model fitting approach consists of in more detail:

1. **Regulator selection:** We selected all genes having GO annotations as transcription factors, kinases or epigenetic factors as potential regulators.

2. **Expression scaling:** As we were mainly interested in finding regulator-target pairs which showed a similar but time-shifted shape of response, we scaled the absolute expression values dividing them by the maximum range of measured expression per gene:

   \[ y_s(t) = \frac{y(t)}{\max_{i=1:T}(y(t_i)) - \min_{i=1:T}(y(t_i))} \]  

   (3.14)

3. **Regulator fitting:** For each of the regulators, we fitted impulse models to their scaled expression profiles by local optimization using a weighted conjugate gradient descent procedure (details in the next paragraph) starting from 100 random parameter initializations. We computed fitting error as squared distance of the impulse curve to the scaled expression values and kept the fit resulting in lowest fitting error as the optimal solution.

4. **Gene-pair fitting:** For each regulator, we computed impulse model fits to genes potentially regulated by it. To this end, we shifted the target gene’s expression profile by between 15 minutes and four hours (in steps of 15 minutes) and then fit an impulse model to the set of data points obtained by merging the regulator’s and the shifted target gene’s scaled expression values. Gene-pair fitting was done using weighted conjugate gradient descent starting from the best parameter set resulting from the regulator-only model fitting in step 3. The time shift leading to the best fitting model (lowest fitting error) was selected as the most likely solution and the corresponding model’s parameters and fitting error used for the subsequent analysis.

Weighted conjugate gradient descent optimization In their paper, Chechik and Koller describe a conjugate gradient descent procedure they use for fitting the impulse model to measured data. We adopted their procedure and extended it by putting weights on the individual measured time points. Weights were computed to reflect the degree to which the two replicate measurements at a given time point \( t \) agree with each other. In detail, weights for the measurements of each gene were computed from the replicate-specific differential expression profiles versus control \( \text{dexp}_r1 \) and \( \text{dexp}_r2 \) as follows:

```matlab
function [weights] = weights_from_range( dexp_r1, dexp_r2 )
% get the bigger and smaller value over the two replicates at each time point
maxvals = max([dexp_r1;dexp_r2]);
minvals = min([dexp_r1;dexp_r2]);

% get the difference between replicates at each time point
diffs = maxvals - minvals;
```
3.2. Modeling responses to abiotic stresses

% get the maximal range of expression over both replicates
maxrange = max(maxvals) - min(minvals);

% compute raw weights
raw_weights = maxrange - diffs;

% normalize raw weights to sum up to the number of time points
weightsum = sum(raw_weights);
weights = numel(raw_weights) * (raw_weights/weightsum);

end

Code 3.1: Weight computation

The resulting weights are positive, reflect the observed difference between replicates at each time point in relation to the range of expression that is spanned over the whole time-course and sum up to the number of time points. Time points showing a small difference between replicates will receive higher weights than time points showing bigger replicate-differences.

We then incorporated the weights \( w \) into the error function \( E \) defined by Chechik and Koller

\[
E = \frac{1}{2} \sum_{i=1:T} (f_\theta(t_i) - y_i)^2
\]

yielding a weighted error function \( E_w \):

\[
E_w = \frac{1}{2} \sum_{i=1:T} (f_\theta(t_i) - y_i \cdot w_i)^2
\]

The estimation of the gradient as the partial derivative of the error function with respect to \( \theta \) was adapted accordingly by replacing

\[
\frac{\partial E}{\partial \theta} = \sum_{i=1:T} \left( f(t_i) - y_i \right) \frac{\partial f(t_i)}{\partial \theta}
\]

with

\[
\frac{\partial E_w}{\partial \theta} = \sum_{i=1:T} \left( f(t_i) - y_i \right) \frac{\partial f(t_i)}{\partial \theta} w_i^2
\]

Finally, we tested whether incorporating information about the replicate agreement at different time points in the form of weights improved parameter estimation. To this end, we set up a test case with a set of known impulse model parameters \( \theta^* \). From this model, we computed the expression \( y^* \) at a set of given time points (0min, 15min, 30min, 1h, 3h, 4h, 6h, 12h and 24h) and then, in 200 repetitions, created two replicate expression vectors by adding Gaussian noise with different standard deviations to the individual time points. A noisy expression vector \( y_{\text{mod}} \) was computed as the mean of the two replicates in each of the 200 repetitions, weights were computed from it as outlined above, and we then fitted two impulse models — one using Chechik’s fitting procedure, the other one our weighted fitting procedure — to \( y_{\text{mod}} \). We evaluated the results of the test case by comparing the squared difference between the true expression data \( y^* \) and the expression \( \hat{y} \) predicted by either of the two impulse model fits. The supplementary material contains
code to set up, carry out and evaluate this test case (see code A.1). The weighted fitting procedure resulted in a small, but significant, prediction improvement over non-weighted fitting \((p = 0.0011, \text{paired one-tailed t-test})\). For data sets having more than two replicate measurements, the estimation of the degree of deviation between replicates should be able to yield more robust time-point weights and thus enable bigger improvements of the fitting procedure over non-weighted fitting.

**Regulator-target pair selection** The fitting procedure outlined above puts more emphasis on fitting the regulators profile closely than on matching the time-shifted target gene profiles. It was designed in that way to yield low fitting error in case a potential target-gene’s time-shifted expression matches the shape of the regulator gene well. We thus used the model fitting error to scan for likely regulator-target pairs.

In order to determine which regulator-target pairs are most likely to represent true regulatory interactions, we devised a selection scheme based on false discovery rate estimates. To this end, we compiled a set of gene-pairs that we did not expect to show true regulatory interactions. This negative set contained all possible pairs of genes with 'regulators' that had no GO annotation indicating them as exerting a regulatory function. We then fit an impulse model to each gene-pair in the negative set using the same procedure as outlined above and evaluated the difference between resulting fitting errors in the negative set and the set of candidate regulator-target pairs in the style of a ROC analysis. To determine a fitting error cutoff yielding an enrichment of hits in the candidate set versus the negative set, we computed the discovery rate \(DR\) in a set of gene-pairs at a given error threshold \(c\) as:

\[
DR_{\text{set}}(c) = \frac{|\{gp \in \text{set} \mid E_w(gp) < c\}|}{|\{gp \in \text{set}\}|}
\]  

We then defined the false discovery odds \(FDO\) between the candidate and the negative set as:

\[
FDO(c) = \frac{DR_{\text{neg}}(c)}{DR_{\text{cand}}(c)}
\]  

Finally, we choose the largest cutoff \(c\) yielding false discovery odds of less than 10% and retained all gene-pairs with fitting errors \(\leq c\). As the set of 'negative' gene-pairs was compiled solely based on a lack of annotations as regulators, this selection procedure can be regarded as conservative, since it cannot be excluded that some of the genes do have regulatory functions but lack corresponding annotations.

**Response network creation** Applying the above described criteria led to a selection of 478 regulator-target pairs between 115 genes in the heat and 106 interactions between 51 genes in the drought data set for the root tissue. For the shoot tissue data, no threshold \(c\) yielding false discovery odds of less than 10% could be found for either the heat or the drought data. The respective sets of selected regulator-target pairs were then visualized as a regulatory network in Cytoscape.
3.2. Modeling responses to abiotic stresses

Predicted regulatory network for drought stress

The inferred regulatory network for the drought root tissue data set is shown in Figure 3.15. It is noteworthy that many of the earliest induced regulators are transcription factors of the AP2/ERF or AP2/DREB family. In total, 8 AP2 transcription factors ERF13, ERF53/AT2G20880, CBF1, CBF2, another DREB1A family factor, DDF1, a DREB4A factor AT5G52020 and DREB5A factors ORA47 and AT1G19210 are contained in the network and 11 of their 27 direct targets have at least one instance of the core DRE-binding motif (sequence RCCGAC) in their promoters (enrichment p-value = 0.01 according to the Athena tool [O'Connor et al., 2005]). The MYB-R2R3-type transcription factor MYB77 has previously been reported to be involved in auxin-signaling and lateral root growth [Shin et al., 2007]. The promoters of the six predicted MYB77 targets all have a MYB-binding motif (sequence MACCWAMC). Due to the frequent occurrence of this motif, however, no significant enrichment was detected. Other early induced predicted regulators were the calcium-dependent kinase CPK32 and the zinc finger protein DOF1. CPK32 has been shown to interact with and phosphorylate the ABA-induced transcription factor ABF4 [Choi et al., 2005] and all three of its predicted target genes have ABRE-binding sites within 400 bp of their transcription start site (enrichment p-
3.2. Modeling responses to abiotic stresses

value = 0.006). Fitting with this potential mode of action, one of the three targets is Late Embryogeneis Abundant 4-5 (LEA4-5), of the group 4 LEA proteins which have been shown to be ABA-inducible and to confer enhanced tolerance to salt and drought stress [Dalal et al., 2009]. The function of DOF1 has not yet been elucidated in detail.

Predicted regulatory network for heat stress

The full inferred regulatory network for the heat stress root tissue data set is too large to be displayed in a figure. Since our main interest was in finding key regulatory mechanisms shaping the heat stress response, we focused on a small sub-network of genes predicted to mediate most of the earliest regulatory control events. Figure 3.16 shows these early regulatory interactions. The zinc-finger transcription factor At3g53820 is first induced in the regulatory cascade shown in panel A. Its average predicted onset time (parameter $t_1$) is about 30 minutes and at full induction, its expression is about 2.5 times higher than in control conditions. It has not yet been implicated in the regulation of heat responses. HSFA7B has an average predicted onset time of about 40 minutes and is highly induced (FC vs control > 200) at the one hour time point. It has previously been

Figure 3.16: Impulse sub-network for early heat regulations occurring in the root tissue: (A) The network spanned by inferred early regulations under heat stress. The horizontal positioning of genes reflects their average onset time according to the respective impulse model fits. Thickness of edges is inversely proportional to the gene-pair impulse fitting error. Assumed time-delays between regulator and target gene expression are noted on the edges. (B) Expression data and fitted impulse models for a selection of regulator-target pairs shown in A.
3.2. Modeling responses to abiotic stresses

shown to be induced by both heat and high light conditions. Under the latter stress, HSFA7B induction is impaired in HSFA1D/E double knockout plants [Nishizawa-Yokoi et al., 2011]. The drought-response transcription factor DREB2B is also induced under heat and high-salinity stress. Its average onset time was estimated at around one hour and it reaches its full induction (FC vs control of 4) at the three hour time point. The expression of HSFA3 is induced later than the expression of other HSFs like e.g. HSFA2 during heat shock. In the heat stress data, it shows only very slight induction after one hour and reaches its full induction (about 8 times higher than under control) at the three hour time point. Its average onset time was predicted at about one and a half hour. Its protein concentration has previously been shown to follow the observed transcriptomic responses to heat stress closely with a later apparition of the protein in comparison to other HSFs and detectable protein levels until late in the stress recovery phase [Schramm et al., 2007]. Schramm et al. also showed that HSFA3 transcription is directly activated by binding of DREB2A and DREB2B to DRE elements in its promoter. The gene At3g45430 encodes a plasma-membrane associated lectin receptor kinase and has not been implicated in heat stress responses previously. It is induced two-fold over control at the one hour time point and has an average onset time of 45 minutes. The gene At2g38340, also known as DREB19 is a member of the DREB subfamily A-2, which also contains DREB2A and DREB2B. It has been shown to be highly responsive to salt, heat and drought conditions and overexpression of DREB19 conveyed increased salt and drought tolerance, while showing no developmental deficiencies or obvious phenotypes under control conditions [Krishnaswamy et al., 2011]. Its expression kinetic is similar to the one of DREB2B, but has a lower induction amplitude (2-fold over control).

Discussion of impulse network results

Generally, the derived impulse networks show a high degree of ambiguity in terms of assignment of regulators to targets (most targets are assigned to multiple possible regulators) and most predicted regulations are predicted to act quickly (with a delay of 15 or 30 minutes). The bias towards short delays is likely due to the fact that most deregulated genes show early responses and that there are few genes first deregulated at the later time points (c.f. Table 3.4). The ambiguity in terms of assigned regulators reflects that transcriptomic data alone — measured at relatively few time points — does not contain enough information to clearly distinguish between true regulatory interactions and gene-pairs showing similar time-shifted expression shapes by chance. The predicted regulatory interactions were, however, enriched for biologically plausible regulator-target pairs as indicated by the presence of transcription factor binding sites in the promoters of predicted targets. Furthermore, the earliest responding genes in the heat network did contain the experimentally verified regulatory interaction between DREB2B and HSFA3 and predicted another gene of the DREB2 family, DREB19 to exert a similar regulatory function on HSFA3. Consequently, we believe that the impulse model is able to predict biologically plausible sets of regulatory interactions playing a role in the responses to heat and drought stresses. The predicted networks have, however, to be regarded as a set of candidate regulatory interactions. By selecting early regulators and their predicted targets, the biologically most interesting regulations can then be tested in small-scale follow-up experiments.
3.3. Synopsis

Experimental evaluation

To test the impulse model predictions, we selected — after discussion with the plant biologists in Prof. Hirt’s group — a set of novel potential regulators to be tested experimentally. From the drought network, we selected the CBL-interacting protein kinase 2 (CIPK2) also known as SNF1-related protein kinase 3.2 (SNRK3.2), the DREB4A factor At5g52020, the ethylene response factor ERF13 and the so-far functionally uncharacterized transcription factor DOF1. From the heat sub-network we selected the three less-well characterized genes At3g53820, At3g45430 and At2g38340/DREB19. Pierre Jacob, a PhD student in Prof. Hirt’s group kindly agreed to obtain knock-out mutants for the above-mentioned candidate proteins and measure the response of their predicted target genes under heat and drought stress conditions. The experimental evaluation is — at the present time — still ongoing. Preliminary data indicates that knock-out plants for DREB4A/At5g52020 under control conditions have an enhanced water content, which would be consistent with the behavior of plants sensing an oncoming drought period.

3.3 Synopsis

Within this project, we have developed data pre-processing protocols specifically tailored towards modeling regulatory mechanisms involved in plant stress responses to abiotic stresses. This included the choice of an appropriate microarray normalization method and gene selection criteria to ensure the selection of robust and relevant portions of data sets measuring the response of a biological system to external stimuli in a time-resolved manner. We applied these protocols to the AtGenExpress abiotic stress data set and then used two modeling approaches in order to predict regulatory networks involved in the response to the measured stresses.

The first model, published by Shiraishi et al. is a state-space model which groups individual genes into clusters and predicts regulatory interactions between them. Once fit to the data, it allows predicting the effect a simultaneous application of multiple stresses would exert on the measured biological system. The regulatory network it predicted to underlie the cold, osmotic and salt stress response data did, however, exhibit very poor consistency with known elements involved in the response to those stresses. This is most likely caused by overfitting of the model due to the high level of complexity inherent in the plants’ stress sensing and response machinery and the limited number of time-points at which the plant stress responses were measured.

Due to these findings, we developed a less complex modeling approach which does not attempt to find a holistic model trying to explain multiple stresses at a time, but instead focuses on predicting individual regulator-target pairs involved in the response mechanisms of a single abiotic stress. The approach is based on the assumption that a given regulator and its potential targets should follow a similar but time-shifted expression kinetic. It makes use of a mathematical model describing impulse-like gene responses developed by [Chechik and Koller, 2009]. We extended the model fitting procedure to include information from experimental replicates and used it to simultaneously fit both the regulator and target kinetics with an assumed time-lag to cancel out deleterious effects due to lacking fitting robustness. We showed that the inclusion of replicate information im-
3.3. Synopsis

proves the model fitting procedure already for data with two replicates. The improvement of this fitting procedure over the one described by Chechik et al. should be substantially larger on data sets with more than two replicate measurements.

We applied our modeling approach to the heat and drought stress portions of the At-GenExpress data set, which measure a transient application of the stress and the plants recovery after stresses have been alleviated and are enriched for impulse-like response patterns. The predicted regulatory networks for the heat and drought stress data were enriched for biologically plausible regulator-target pairs and experimental testing of the most promising candidates is currently being carried out.
3.3. Synopsis
Chapter 4

Single-cell transcriptomic signatures elicited by *Shigella* infection

4.1 Background

This project was carried out in collaboration with the 'Dynamics of Host-Pathogen Interactions' unit of the Pasteur institute in Paris headed by Dr. Jost Enninga. His group’s focus is on developing novel approaches to investigate the interactions between pathogens and their hosts in single cells in real time. To this end, they develop novel assays that allow monitoring pathogen invasion; simultaneously tracking a large number of parameters involved in the invasion processes. With the help of those methods they strive to correlate the pathogen behavior and its intracellular localization to the host responses it induces.

In the scope of the project presented here, Juliane Lippmann, a post-doctoral researcher in Jost Enninga’s group, established the screening protocol, did the experimental work and data collection, was closely involved in developing the computational analysis pipeline and supplied the biological expertise needed in interpreting the analysis results. We are currently preparing a manuscript presenting our findings for submission.

4.1.1 Host-pathogen interactions

Pathogen-sensing and immune response

Host cell surfaces are constantly colonized by a large variety of microbes, however, occasionally infection by both extra- and intracellular pathogens interferes with the host cell signaling homeostasis. To discriminate between non-pathogenic settings and harmful threats, the host cellular immune system senses so-called patterns of pathogenesis [Vance et al., 2009]. This involves the distinction between extracellular and intracellular microbe- and/or non-microbial danger-associated molecular patterns (MAMPs or DAMPs) by a dual system of pattern-recognition receptors (PRRs). Those are expressed at host membrane compartments, such as the Toll-like receptors (TLRs), and in the cytosolic compartment, including the Nod-like (NLR) receptors [Takeuchi and Akira, 2010]. PRRs
4.1. Background

recognize bacterial cell wall components (e.g. lipopolysaccharide (LPS) or peptidoglycan (PG)) as well as a wide range of DAMPS (e.g. extracellular ATP, K+ efflux). PRR stimulation activates a number of signaling pathways involving the signaling molecules IKK, RIPK2 and MAPK and subsequently the transcription factors NFkB, cJun and IRF3 leading to the induction of a range of cytokines and chemokines (e.g. IL-8, IL-6, CXCL-2, IFNs) and antimicrobial peptides (e.g. HBD3). Using single-cell imaging methodologies, it has recently been described that inflammatory responses to intracellular infections might also emanate from non-infected neighboring cells ([Kasper et al., 2010]; [Ablasser et al., 2013]) as a host cell-cell communication strategy to circumvent pathogen-triggered immune suppression.

In addition to MAMP/DAMP recognition, the cell is also able to sense perturbations of core cellular processes, including actin cytoskeleton disruptions, the cytosolic access of microbes or microbial molecules, intracellular growth and spread, changes of nutrient levels and different cell stresses, which lead to a number of transcriptional and post-translational host immune responses. Specifically, infections with intracellular bacterial pathogens have been shown to trigger endoplasmic reticulum (ER) and integrated stress pathways ([Chakrabarti et al., 2012]; [Tattoli et al., 2012]). Both pathways converge at the translation inhibitor molecule eIF2α, which blocks global translation of most mRNAs, while preferentially translating ATF4 and ATF3, two major transcription factors of the integrated stress response. The ATF3-dependent transcriptional signature includes regulation of genes involved in stress, apoptosis and cell cycle regulation, such as the downstream targets CHOP and CHAC1, the stress sensor growth arrest and DNA damage 45α (GADD45α), the cell cycle regulator kinase CDKN1A/p21 or genes of DNA repair and detoxification including ERCC2, XRCC5 and CRY1.

*Shigella flexneri* infection

*Shigella flexneri* is a gram-negative, intracellular bacterial pathogen that invades intestinal epithelial cells from the basolateral side after hijacking and killing intestinal macrophages by an inflammatory cell death, called pyroptosis. *Shigella* pathogenicity originates from its virulence-plasmid encoded type three secretion system (T3SS), by which a set of effector molecules is injected into the host cytosol that interfere with diverse host cell signaling pathways during bacterial invasion of host cells ([Buchrieser et al., 2000]; [Sansonetti, 2001]). *Shigella* entry into epithelial cells unfolds in a two-step process (c.f. Figure 4.1). In a first, membrane-bound, step, it adheres to the host cell membrane initiating actin foci formation and internalization into a pathogen-containing vacuole ([High et al., 1992]; [Lafont et al., 2002]; [Monnier et al., 1999]; [Nhieu et al., 1999]). Then, *Shigella* triggers vacuolar rupture, it replicates in the cytoplasm, and moves intra- and intercellular by actin-polymerization ([Ray et al., 2010]). During the invasion, it subverts host immune responses including proinflammatory cytokine signaling and autophagy, nevertheless it is unclear which host pathways are induced at the successive infection steps ([Shere et al., 1997]; [Nhieu et al., 1997]; [Philpott et al., 2001]).

**Shigella effectors** T3SS-mediated injection of bacterial effectors occurs in two waves dependent on two transcriptional activators, VirB and MxiE, respectively ([Mavris et al., 2002]; [Gall et al., 2005]). The effector OspF, which is secreted in a VirB- and MxiE-
4.1. Background

Figure 4.1: Schematic representation of *Shigella flexneri* infecting an epithelial cell: (1) *Shigella* attaches to the outer membrane and injects its effectors using a Type-III Secretion System (T3SS). The effectors lead to actin remodelling resulting in foci formation, which enables uptake of the bacteria into a pathogen-containing vacuole (2). *Shigella* triggers vacuole rupture and enters the cytosol of the host cell. Through actin-polymerization *Shigella* is capable of intra- and inter-cellular movement (3) and the infection spreads to neighboring cells (4).

dependent manner, is a dual phosphatase that interferes with MAPK signaling pathways by irreversibly dephosphorylating p38 and ERK, thereby reducing levels of proinflammatory cytokines, such as IL-8 ([Arbibe et al., 2007]; [Li et al., 2007]). The primary effector IpgD is a inositol 4-phosphatase that specifically dephosphorylates phosphatidylinositol 4,5-bisphosphate (PtdIns-(4,5)P2) into PtdIns-5-monophosphate (PtdIns(5)P) and thus impacts cytoskeletal rearrangements, dampens host inflammation through extracellular ATP, and activates the PI3K/Akt antiapoptotic signaling pathways [Niebuhr et al., 2000]. The secondary, MxiE-dependent effector OspG blocks NFkB-dependent immune signaling by preventing degradation of IkB [Kim et al., 2005].

**Aim of the project**

Both the dynamics of bacterial subcellular localization during the successive stages of host cell infection and the effector-triggered pathogenesis might influence and differentially regulate the induction of host immune response signaling cascades. Gene expression analysis of immune responses so far has been done either by measuring averaged expression of whole cell populations or by using biochemical methodologies, which result in limited spatio-temporal resolution. Consequently, an integrative picture containing information on the impact of subcellular localization changes on transcriptional immune responses has
4.1. Background

been only insufficiently drawn. The aim of this project was to identify host immune response gene expression signatures that correspond to the distinct stages of *Shigella* invasion into epithelial cells. Therefore, three stages of bacterial infection were determined that potentially activate distinct immune response signaling pathways (c.f. Figure 4.2): (i) neighboring, non-infected cells (NI), infected cells containing bacteria (ii) in membrane-bound or vacuolar localization (vac) and (iii) in the cytosolic compartment (cyt).

![Figure 4.2: Distinct stages of *Shigella* infection](image)

**4.1.2 Experimental Setup**

In order to assess stage-specific transcriptional signatures corresponding to individual stages of bacterial infection in small cell populations and single cells, a workflow was set up, which combines FACS-based cell sorting of distinct infection stages with multiplex qPCR analysis. To achieve this, a CCF4/β-lactamase FRET assay, previously developed for bacterial vacuolar rupture studies at high temporal resolution ([Ray et al., 2010], [Nothelfer et al., 2011]; [Simeone et al., 2012]), was employed (Figure 4.3A). In this assay, a cephalosporin-based FRET probe (CCF4) is loaded onto cells and trapped in the host cytosol. Enzymatic cleavage of CCF4 is induced only upon cytosolic access of β-lactamase (bla), e.g. upon cytosolic escape of bla-expressing bacteria, resulting in a fluorescence switch from 535 nm (green channel) to 450 nm (blue channel). To further distinguish infected and non-infected cells, dsRed-fluorescent *Shigella* are used for infection. Distinct combinations of green, red and blue fluorescence are used in a FACS gating strategy capable of distinguishing between non-infected cells, cells with *Shigella* attached to the outer membrane or inside a vacuole and cells with *Shigella* localized in the cytosol (Figure 4.3B).
4.1. Background

The assay described above was used to sort small bulk cell populations (20 cells) as well as single cells for non-treated control samples and the respective infection stages (Figure 4.3C). For each of those four sample groups (control (CO), NI, vac and cyt), 20 single cells and four bulk cell samples were sorted totaling 96 samples per infection. RNA from the collected samples was further processed for multiplex qPCR analysis using microfluidic devices (BioMark, Fluidigm). For each infection, the expression of 96 genes over the collected 96 samples was measured on a single BioMark array. The 96 genes measured on the BioMark arrays were selected in order to collect maximal information on the host immune response. Contained were genes involved in shaping the early host immune response and genes likely to be altered on the transcriptional level during Shigella infection. This included genes involved in inflammation and PRR signaling, apoptosis, cellular stress response and repair (see Table A.1 in the appendix). Furthermore, two reference genes, B2M and PGK1, were included for normalization of bulk cell sample CT values. A detailed description of the experimental protocols for infection, cell sorting, sample collection and pre-processing as well as for transcriptomic measurements, conceived and carried out by Juliane Lippmann and Jost Enninga, is included in the appendix as supplementary material (see section A.3).

The main data set analyzed in this project consisted of three replicate experiments. Each experiment consisted of six BioMark arrays, each individual one measuring the response to infection by a specific Shigella strain (infections were done with either wild type (WT) M90T or mutants deficient for one of the five Shigella effectors OspF, OspG, IpgD, MxiE and MxiD). In total, our collaborators collected over 40 Biomark arrays, measuring — apart from the main data set mentioned above — a number of different experimental conditions like, e.g. different infection parameters (multiplicity of infection, time after infection), the reaction to simplified stimuli like LPS and the effect knock-downs of central host immune receptor and signaling components have on the immune response. To deal with this large collection of data, we developed an analysis pipeline encompassing the most important steps in data processing and facilitating interpretation of results.
Figure 4.3: Experimental workflow developed for the analysis of host immune responses to different stages of *Shigella* infection: (A) The CCF4/β-lactamase FRET assay (B) Distinguished stages of bacterial infection (C) Cell sorting and transcriptomic measurement of host immune responses. A detailed description is given in the accompanying text.
4.2 Data analysis pipeline

Figure 4.4 shows an overview of the data analysis pipeline we developed for the collected BioMark data. The pipeline is implemented in MATLAB and makes use of custom R scripts for visualization and single-cell-specific analysis steps. For each analyzed qPCR array, CT values obtained with the BioMark Real-time PCR Analysis software (Fluidigm) are read-in from the proprietary Fluidigm output format and filtered according to the presence/absence calls made by the Fluidigm software.

In the following, we will describe the individual steps of our analysis pipeline and, where parameters needed setting, state the values used for analysis of the main data set described above.

Figure 4.4: Overview of the analysis pipeline: Data read-in directly from the proprietary Fluidigm output file format are subjected to quality control, which filters out individual samples and genes of poor quality. Data passing the quality control is subjected to normalization and analysis procedures e.g. in the form of detection of differential expression between experimental conditions and principal component analysis. Data from individual arrays is tested for compatibility and can be pooled into a single data set allowing comparison between samples measured on different arrays. The pipeline furthermore creates commonly used analysis plots like e.g. box- or violin-plots showing the expression of a given gene under different measured conditions.
4.2. Data analysis pipeline

4.2.1 Pipeline components

**Quality Control**  Arrays with very few valid measurements can easily be identified by visualizing the measured expression as heat maps (e.g. the second array in Figure 4.4). The design of the Biomark microfluidic arrays as well as sample and primer preparation can, however, also lead to individual rows (samples) or columns (genes) showing abnormally low rates of expression. To filter those out, we devised the following selection criteria: genes showing CT values lower than a user-defined threshold in less than \( n \) samples, as well as samples showing CT values lower than the same threshold in less than \( n \) genes were discarded from the analysis. For the analysis of our main data set, we set the expression threshold to a CT value of 40 and \( n \) to 6 out of 96.

**Data normalization**  Analysis of bulk cell qPCR data requires a normalization procedure to correct for differing global levels of mRNA between individual samples. Typically, this is achieved through computing so-called \( \Delta \)CT values by subtracting the CT values measured for a reference gene from the CT values measured for the genes of interest. The right choice of reference gene is crucial for further data analysis. Our pipeline creates plots showing the average expression level and variance between samples for all measured reference genes as well as for the mean expression of all possible combinations of reference genes. The user can then select the reference gene or combination of reference genes showing the least variance over samples as reference for \( \Delta \)CT value computation. In the case of the bulk cell portion of the data presented here, the mean expression of PGK1 and B2M showed least variance over samples and was thus selected as reference for normalization. Single cell data does not allow normalization via a reference gene since expression occurs in stochastic patterns of transcriptional bursts, which are not necessarily synchronized between individual genes [Raj et al., 2006]. We thus performed filtering of potentially non-specific single cell expression measurements (see section 4.2.2 for details) and used non-normalized single cell expression measurements for the subsequent analyses.

**Pooling individual arrays**  In order to detect technical variations between individual arrays, within one experiment, all arrays were loaded with cDNA of the same bulk and single cell control samples. Our pipeline computes array-specific deviations from the average expression over all arrays and shows them in a scatter plot (shown on the bottom-right of Figure 4.4). For two of the three replicate experiments, all arrays showed highly comparable control sample expressions with deviations lower than \( \pm 1 \) CT for reasonably well-expressed genes. The third experiment showed significantly larger technical variability. Further evaluation of the two experiments showing low technical variability showed that their global distributions of measured bulk cell CT values were comparable and that principal component analysis of the pooled data did not separate the samples into groups corresponding to the two experiments. We thus decided to evaluate those two replicate experiments individually as well as in a pooled manner to increase sample size.

**Test for differential expression between infection stages**  The small sample size of the bulk cell data (up to 8 measurements per infection stage after pooling two experiments) did not allow making distributional assumptions needed to apply parametric tests like
4.2. Data analysis pipeline

e.g. the t-test. We thus used a non-parametric Mann-Whitney-U test for the detection of differential expression between individual infection stages. The test for differential expression we applied to the single cell data is described in more detail in section 4.2.2.

**Analysis and Visualization** On data with small sample sizes, relying entirely on statistical tests is not advisable. Consequently, proper visualization of the data showing the measured expression in each individual sample is an important feature for data analysis. Our pipeline produces box plots showing, for each gene, the expression of all samples, grouped by infection stage, along with a table detailing the results of the differential expression test between infection stages (for an example see Figure 4.5A). Since the single cell data encompasses more individual measurements and analyzing expression distribution patterns is of central interest, we visualized single cell data using violin plots (Figure 4.5B).

Our pipeline additionally carries out principle component analysis (PCA) for both bulk and single cell samples and creates corresponding PCA plots. PCA is a mathematical transformation that allows representation of high-dimensional data sets (e.g. 96 dimensions corresponding to information from 96 genes) in low-dimensional space (e.g. 2D or 3D), while retaining as much of the information contained in the data as possible. Thus, the relative distance of data points gives information on the similarity of transcriptional profiles between individual samples.

Genes tested as differentially expressed between any of the studied sample groups are furthermore visualized in heat maps displaying their average expression within sample groups and thus yielding an overview of gene expression patterns characteristic for the individual sample types.
4.2. Data analysis pipeline

Figure 4.5: Data visualization example: (A) A box plot showing the bulk cell expression of the chemokine CXCL-2 in the four sample types measured along with the differential expression test results. (B) Expression of CXCL-2 in the single cell data visualized by violin plots. Its mean expression over single cells reproduces the expression pattern observed in bulk cell populations. The violin plots show, however, that its expression distribution in non-infected cells shows a larger spread between individual cells than in either control or infected cells. The heatmap shows results of a test for differential expression in single cell data explained in more detail in section 4.2.2.
4.2. Data analysis pipeline

4.2.2 Single cell data analysis

CT cutoff evaluation

Although single cell analysis has substantially advanced in recent years owing to great technological improvements, existing techniques to measure single cell gene expression are still hampered by lack of accuracy in expression measurements for lowly expressed genes. In order to eliminate such effects, we examined the concordance between bulk cell and their corresponding single cell CT values. To this end, we computed in-silico pooled expression ($CT_{SC}$) over individual single cell samples of the same sample type and compared them to the average CT of the corresponding bulk cell samples ($CT_{BC}$). In detail, $CT_{SC}$ was computed by transforming single cell CT values to expression threshold (ET) values, summing up individual single cell ETs, normalizing the summed-up expression to the number of cells contained in the bulk cell samples and transforming them back to CT values:

$$ET(i) = \max\{0, CT_{\text{max}} - CT(i)\}$$
$$ET_{\text{lin}}(i) = 2^{ET(i)}$$
$$ET_{\text{sum}} = \frac{n_{bc}}{n_{sc}} \sum_{i=1}^{n_{sc}} ET_{\text{lin}}(i)$$
$$ET_{SC} = \log_2 ET_{\text{sum}}$$
$$CT_{SC} = CT_{\text{max}} - ET_{SC}$$

In the above equations $n_{bc} = 20$ (the number of cells contained in a bulk cell population) and $n_{sc}$ indicates the number of single cells in a given sample group that have passed the quality control filter.

As can be seen from the equations above, the computation of $CT_{SC}$ involves setting a threshold $CT_{\text{max}}$, above which all expression measurements are regarded as non-expressed. For each experiment, we thus evaluated different choices for $CT_{\text{max}}$ and selected the one maximizing the concordance between the mean over bulk cell CTs below $CT_{\text{max}}$ and $CT_{SC}$. The results of the $CT_{\text{max}}$ selection procedure applied to pooled data from the two congruent experiments are shown in Figure 4.6A. It shows the values of two concordance measures as well as the proportion of retained data dependent on the choice of $CT_{\text{max}}$ ranging from 20 to 40. Setting $CT_{\text{max}}$ to 23 yielded optimal concordance according to both measures and retained at least one single and bulk cell measurement for over 90% of the measured gene-sample type pairs. The same CT cutoff was identified when analyzing the two experiments separately. Figure 4.6B shows that applying a CT cutoff renders the single and bulk cell data more comparable by removing noisy data points. We thus set $CT_{\text{max}}$ to 23 and based our further single cell analysis on ET values computed using this cutoff.

Differential expression test according to [McDavid et al., 2013]

To test for differential expression of a gene between infection stages, we used the likelihood ratio test described in [McDavid et al., 2013]. It simultaneously tests for differences in the mean expression $\mu$ over all cells and in the proportion $\pi$ of single cells expressing a given gene at all. The test is based on comparing the goodness-of-fit of two alternative models to the measured data. The null model is based on the assumption that the two experimental
4.2. Data analysis pipeline

Figure 4.6: Evaluation of possible expression thresholds $CT_{\text{max}}$: (A) Shown are the mean absolute difference between $CT_{\text{SC}}$ and $CT_{\text{BC}}$ (in red) as well as their root mean squared deviation (RMSD, in blue) and the proportion of data retained (in green) after applying different expression cutoffs. (B) Scatter plot of bulk cell data ($CT_{\text{BC}}$) versus in-silico pooled single cell data ($CT_{\text{SC}}$) before (grey dots) and after (colored squares) applying a CT cutoff of 23.

units to be compared are identical in terms of the parameters $\mu$ and $\pi$. The alternative model allows separate parameters for the two experimental units to be compared (in our case two stages of infection). Goodness-of-fit of the two models is individually evaluated by the likelihood function:

$$L(\theta | y, v) = \prod_k \pi_k^{n_k} (1 - \pi_k)^{1-n_k} \prod_{i \in S_k} g(y_{i,k} | \mu_k, \sigma^2)$$

where $y$ is the vector of measured ET values for a given gene across the two groups, $v$ is a binary vector indicating presence of expression, $\theta$ is the vector of parameters to be fitted to the data, $S_k$ is the set of cells expressing the gene in group $k$ (i.e. $S_k = \{i | v_{i,k} = 1\}$), $n_k = \sum_i v_{i,k}$ is the number of cells expressing the gene in group $k$ and $g$ is the density function of the log-normal distribution with parameters $\mu_k$ and $\sigma^2$. The likelihood ratio test statistic is then defined as the ratio of the likelihoods of the null and alternative model and can be shown to asymptotically follow a $\chi^2$-distribution with two degrees of freedom under the null hypothesis \cite{Wilks_1938}, which allows the computation of p-values for differential expression. The likelihood ratio is furthermore decomposable into the effect mediated by differences in terms of the mean expression $\mu$ and the effect mediated by different proportions of expressing cells $\pi$. Such a decomposition allows evaluation of the type of differential expression between two sample groups. In the violin plots created by our analysis pipeline this is indicated by a ‘c’ if the continuous part ($\mu$) of the test is dominant, by a ‘p’ in case the proportional part ($\pi$) of the test is more significant and by a ‘b’ if both of them contribute equally to the test result (c.f. Figure 4.5B).
4.2. Data analysis pipeline

Test for bimodality

Single cell data allows, in contrast to population-level data, for analysis of variations in the degree of expression over individual cells. Whereas in population-based data, the average expression and changes thereof are analyzed, in single cell data the distribution of expression values can be of high interest. A change from a unimodal distribution — with all single cells expressing a given gene at similar levels — in control samples to a bimodal expression upon infection — where a subset of single cells react stronger to the infection than others — indicates heterogeneity within the infected cells and merits closer investigation. We thus tested the single cell data for genes showing bimodal expression patterns using a measure called bimodality index (BI). First defined in Wang et al. [2009], BI is a criterion to identify and rank bimodal signatures from gene expression data. It is assumed that the distribution of a gene with bimodal expression can be expressed as a mixture of two normal distributions with means $\mu_1$ and $\mu_2$ and identical standard deviation $\sigma$. The standardized distance $\delta$ between the two populations is defined as:

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma} \quad (4.3)$$

The computation of BI requires classification of the data into two normal distributions following the assumptions mentioned above, which can for example be achieved using the R package mclust [Fraley and Raftery, 1999]. Given such a classification, BI is then computed by correcting the $\delta$ score for the distribution of data points into the two classes:

$$BI = \sqrt{\pi(1-\pi)} \ast \delta \quad (4.4)$$

with $\pi$ indicating the proportion of data points assigned to the first class. The BI can be used as statistic in a test based on a null hypothesis assuming $\delta = 0$. Wang et al. recommend BI = 1.1 as a cutoff to select bimodally distributed genes when applied to data sets containing at least 100 samples. This cutoff does, however, not translate to cases with fewer samples. In simulated data fulfilling the null hypothesis of unimodality (data drawn from a single normal distribution with varying standard deviations), we observed a strong dependency of BI values on the number of measured single cell expression values $n$. We thus repeatedly drew $n$ random samples from normal distributions with $n$ ranging from 6 to 40 (10,000 repetitions for each value of $n$). Based on these random samples, we computed BIs and estimated — separately for each sample size $n$ — the quantiles of its empirical distribution function. Based on these quantiles, BI cutoffs corresponding to different test significance levels can be defined for each sample size $n$. Sample groups where less than 6 single cells showed expression of a given gene were deemed non-conclusive and thus discarded from the bimodality analysis.

Pathway correlation analysis

Genes functioning in the same molecular pathway are often thought to be induced or repressed concomitantly due to being controlled by common regulatory elements. To test whether bacterial invasion localization- and/or effector-dependently affects host cell signaling circuits, we carried out pathway-based correlation analyses.
4.2. Data analysis pipeline

**Gene-pair correlations**  Gene-pair correlation analysis was performed using Spearman’s rank correlation coefficients, which were computed for all possible pairs of genes on ET values of 20 single cells separately for each condition and experiment. To assess significance of the correlation coefficients, a background set of 10,000 correlations computed over randomly permuted orderings of the 20 single cells was constructed. An empirical p-value, which indicates significance of the absolute measured correlation, was calculated as the proportion of random permutations that yielded a correlation at least as extreme in absolute value as the measured one. Assuming independence of experimental repeats, a combined p-value \( p_{\text{comb}} \) was calculated by multiplying permutation-based p-values of individual experiments. If disagreement on the sign of correlation between experiments was observed, the experiment with the least significant p-value was discarded.

**Within-pathway correlations** To gain a broader and more robust view of global correlations, we examined the degree of correlation between genes assigned to common molecular pathways. Genes were assigned to one or more of the following pathways manually, after review of significant literature: pro-inflammatory genes, apoptosis-related genes and genes involved in the reaction to endoplasmic reticulum (ER) stress. A full list of the pathways and their associated genes can be found in the appendix (section [A.3]). The degree of correlation within a pathway was measured by a score \( \text{corr}_{\text{sum}} \) computed as the mean of the negative log-transformed \( p_{\text{comb}} \) values over all gene pairs within the pathway. To correct for an observed expression-level bias of the score (highly expressed gene sets tend to show higher \( \text{corr}_{\text{sum}} \) scores than sets comprised of lowly expressed genes), the 96 measured genes were classified into 5 bins of approximately equal size according to their average expression levels. Per studied pathway and sample type, 50 separate p-value computations were run. Per run, 200 pseudo-pathways with the same number of gene pairs and expression-bin compositions than the actual pathway were randomly drawn. Then, \( \text{corr}_{\text{sum}} \) was computed and the number of pseudo-pathways with \( \text{corr}_{\text{sum}} \) at least as extreme as observed in the actual pathway was counted and divided by the number of pseudo-pathways tested. The mean \( \mu \) and standard deviation \( \sigma \) of the 50 resulting pathway p-values was used to define a score measuring significant enrichment of correlations within a pathway. It ranges from 0 to 3 and was constructed by consecutively applying the following rules: A score of 3 was assigned to pathways for which \( \mu + \sigma \) was below a significance level \( \alpha \) of 0.05. A score of 2 was assigned to pathways, which had a mean p-value \( \mu \) below 0.05, a score of 1 to pathways with \( \mu - \sigma \) below 0.05 and a score of 0 otherwise. Scores for significant depletion of correlation in comparison to the pseudo-pathways were analogously computed on inverted p-values \( p_{\text{inv}} = 1 - p \) and were assigned scores from -3 to 0.

**Pathway correlation changes during infection** To test for significant changes in the degree of correlation between two experimental conditions (e.g. two infection stages or the same infection stage under two bacterial strains), an extension of the significance test discussed above was applied. For a given pathway, the difference in \( \text{corr}_{\text{sum}} \) scores between the two experimental conditions was computed and pairs of randomly drawn pseudo-pathways from the same two conditions were used to estimate a background distribution of expected differences in \( \text{corr}_{\text{sum}} \). The p-value computation scheme described in the previous
4.2. Data analysis pipeline

paragraph was carried out on the differences of corr_{sun} and significance of correlation increase or decrease of a pathway from one condition to another was scored on a range of 3 to -3 analogously.

Exploratory analyses

The analysis of correlation structures and changes during infection relies heavily on the classification of genes into biological processes. To get a less biased picture of correlation structures present in the data and how they change during the infection by *Shigella*, we created data representations allowing the exploratory analysis of the data set at hand.

**Correlation networks** We developed R scripts automatically visualizing all gene-pairs with a significant correlation according to a cutoff set on $p_{comb}$ as a Cytoscape [Cline et al., 2007] network. In this network each correlation is visualized as an edge, with edge color and edge width representing the sign and degree of significance, respectively. Individual genes are represented as nodes and the measured expression over all single cells of a given sample type are visualized as a radar plot graphic displayed on the gene node. During our analyses, we found that this visualization greatly facilitates interpretation of observed correlation structures (by e.g. looking at highly correlated clusters of genes in the network) and their potential causes (by closer examination of gene expression patterns over single cells).

**Cell cycle phase prediction** Cyclins are proteins that are involved in control of the cell cycle progression. Their expression levels oscillate over the course of the cell cycle and they have been shown to bind to and activate cyclin-dependent kinases (CDKs). The CDKs, once activated, phosphorylate specific targets and hence initiate the processes needed to advance to the next cell cycle phase. Figure 4.7A shows a schematic representation of expression variations human cyclins A, B, D and E undergo during one cell cycle.

Since the relative cyclin expression levels are closely related to the cell cycle phase a cell is in, they can be used for estimating cell cycle phases of individual single cells. To this end, we devised a simple set of rules, classifying cells into either of the cell cycle phases $G_1$, S, $G_2$ and M according to comparisons of relative Cyclin A, B and E levels normalized to their maximal expression over all single cells on an array. Figure 4.7B shows the average relative cyclin expression of single cells grouped by their predicted cell cycle phase. Our classification scheme was able to closely reproduce the expected expression patterns of cyclins over the individual cell cycle phases. We then proceeded to analyze how the proportion of cells classified into the different cell cycle phases changes during the infection by different *Shigella* strains. To this end, we computed the proportion of single cells in each sample group classified as being in a given cell cycle phase. For each of the sample groups exposed to bacteria (NI, VAC and CYT), we compared these proportions to the proportion measured in the control condition and computed p-values by assuming a binomial distribution with $p$ set to the proportion of cells predicted to be in a given phase in the control condition. We then applied a Benjamini-Hochberg multiple-testing
4.2. Data analysis pipeline

correction to the obtained p-values to identify cell cycle phases significantly enriched or depleted during infection in comparison to the control condition.

Figure 4.7: Cell cycle phase classification: (A) Changes in the relative level of expression of individual cyclins determines cell cycle phase progressions [Source: Wikipedia page on Cyclins, http://en.wikipedia.org/wiki/Cyclin]. (B) Applying a simple classification scheme based on relative levels of cyclin expression closely reproduces the expected patterns of cyclin A, B and E expression.
4.3 Bulk cell results

We applied our analysis pipeline to the pooled data of the two congruent experimental repeats. The resulting plots and statistics as well as further in-depth analyses led to the following key findings:

4.3.1 Host immune responses to distinct stages of *Shigella* infection are distinguishable

In order to identify transcriptional signatures of individual infection stages, we first identified genes showing significant expression differences between any of the individual sample groups recorded during *Shigella* WT infection according to a Mann-Whitney-U test at $\alpha = 0.05$. 41 of the measured 96 genes were deemed differentially expressed and a PCA plot based on these 41 genes showed that the individual sample groups were largely separable (Figure 4.8A). Furthermore, selecting gene sets differentially expressed between control and non-infected cells (20 genes) and vacuolar and cytosolic (15 genes) infection stages, respectively, lead to clear separations of the samples in PCA plots (see Figure A.6 in the appendix). Plotting the average expression per sample group of those 41 genes as a heatmap (Figure 4.8B) revealed that (i) only few genes were down-regulated and most genes induced over the course of infection, that (ii) signatures of non-infected, neighboring cells partly resemble the early (vacuolar) stage of infection and that (iii) the evolution of transcriptional signatures of selected genes dynamically changed during successive stages of bacterial infection as the expression of individual genes either remained unchanged, was significantly altered or was exclusively deregulated at the cytosolic stage.

The established workflow thus enables the dissection of the host transcriptional immune response into the responses to different stages of the infection. This allows the definition of distinct transcriptional signatures corresponding to the vacuolar and the cytosolic stage of bacterial infection as well as to signaling from non-infected cells.

4.3.2 *Shigella* induces a complex T3SS-dependent, but effector-independent bystander response

Activation of inflammatory signaling in neighboring cells, referred to as bystander effect, has been reported for bacterial and viral infections ([Kasper et al., 2010]; [Ablasser et al., 2013]). According to those studies, bystander activation is induced via small signaling molecules, such as the second messenger cGAMP, which are transferred through gap junctions and lead to nuclear translocation of transcription factors. In our study, we used *Shigella* WT-infected HeLa cells, which do not form gap junctions. The observed transcriptional response signatures of non-infected cells did, however, indicate that bystander activation can also be achieved via alternative pathways not involving gap junctions. Thus, we aimed at further characterizing the *Shigella*-induced bystander effect and investigated the impact of different bacterial effectors.

To this end, we selected all genes showing significant change between control and non-infected populations in any of the investigated strains (Mann-Whitney-U test at $\alpha = 0.05$). This identified 23 genes, all of which were up-regulated in bystander cells compared to the
4.3. Bulk cell results

Figure 4.8: Identification of localization-dependent transcriptional signatures: (A) PCA based on genes differentially expressed between any of the sample groups shows that localization-dependent transcriptional signatures are distinguishable (B) Heatmap of genes identified to contribute to gene expression signatures of bulk cells at indicated stages of bacterial infection with *S. flexneri* WT (NI, vac, cyt) and non-treated control (co). Shown are median ∆CT values transformed into row Z-scores between +2 (high expression, yellow) and -2 (low expression, blue)

untreated control (Figure 4.9A). The induction of those genes was highly similar in non-infected cells treated with all tested bacterial strains except for a T3SS-deficient mutant, Δ*mxiD*, which is unable to invade epithelial cells [Allaoui et al., 1993] and did not prompt gene induction. These findings suggest that transcriptional activation in bystander cells is dependent on the bacterial T3SS and that intracellular infection is required before a transcriptional host response in neighboring cells is induced. The bystander response comprised transcriptional induction of three distinct families: (i) proinflammatory genes, including a number of chemokines (such as CXCL-2, Figure 4.9B) as well as components of the proinflammatory signaling pathway (e.g. NFKBIA, RIPK2, cJun), (ii) anti-apoptosis genes (e.g. Birc3, Figure 4.9C) and (iii) genes of the cellular stress response, including the major transcription factors ATF4 (Figure 4.9D) and ATF3 or Gadd45a.

Taken together, the host immune response to pathogenic *Shigella* distinguishes from that to non-pathogenic strains by activation of a bystander response in non-infected neighboring cells. Activation of the bystander response is dependent on the bacterial T3SS, but independent of the bacterial effectors OspF, OspG, MxiE and IpgD and manifests in transcriptional activation of inflammation, anti-apoptosis and the cellular stress response.
4.3. Bulk cell results

Figure 4.9: The bystander response signature: (A) Heatmap showing the expression of the 23 genes comprising the bystander response in non-infected cells extracted from populations infected with either of the *Shigella* strains M90T (WT), $\Delta ospF$, $\Delta ospG$, $\Delta mxiE$, $\Delta ipgD$ or $\Delta mxiD$ (B) Box plot of a representative inflammatory response gene CXCL-2 (C) Box plot of a representative anti-apoptotic response gene Birc3 (D) Box plot of the major cellular stress-associated transcription factor ATF4.
4.3. Bulk cell results

4.3.3 Five main expression profiles shape the immune response to *Shigella* infection

To better characterize the complex host responses and identify how they are regulated dependent on the bacterial subcellular localization, we classified gene responses into major response patterns. To this end, for each gene we scored differential expression between individual sample groups as follows: -1 for a significant down-regulation from the first group to the second, +1 for significant up-regulation and 0 if no significant change had been detected at $\alpha = 0.05$ by a Mann-Whitney-U test. Based on those scores we defined five response profiles using the following decision procedure:

```python
if no significant regulation:
    exclude
if co->ni == 1
    if vac->cyt == -1:
        profile4
    else:
        profile1
else:
    ## no bystander effect
    if co->cyt == 1 and ...
        co->vac == 0 and ...
        co->ni == 0:
            profile3
    else:
        if co->vac == -1 or ...
            co->cyt == -1 or ...
            vac->cyt == -1:
                profile5
        else:
            profile2

Code 4.1: Profile classification scheme
```

The result of applying this classification procedure to the data collected from WT infection experiments is shown in Figure [1,10]. Profile 1 and 2 combine genes, whose expression was induced early upon and remained up-regulated during the course of intracellular infection even after subcellular localization change of the bacteria to the host cytosol. In contrast, genes of profile 3 were slightly, but significantly, up-regulated only upon cytosolic access, indicating that induction of these genes correlated with the compartmental change of intracellular bacteria to the host cytosol. Profile 4 comprised genes that were initially induced at the vacuolar stage, however, were ultimately down-regulated in the course of infection. Finally, down-regulation of genes of profile 5 occurred mainly upon cytosolic bacterial localization. Based on these profiles and Ingenuity pathway analysis, we further explored the localization-dependent transcriptional signatures.
4.3. Bulk cell results

Figure 4.10: Genes classified into distinct expression profiles: Expression of genes classified into each of the 5 profiles and a prototype of each profile are shown in the top row. For each profile, the expression of a representative gene is shown as a box plot in the bottom row.

Vacuolar response  Genes from profile 1, 2 and 4 are part of the overall vacuolar signature of *Shigella* WT-infected cells. This signature comprised transcriptional activation of proinflammatory cytokines and chemokines (e.g. IL-8, IL-6, CXCL-2, CCL-2, IL-18, IFNβ and TNFα), indicating activation of PRR-stimulated signaling via RIPK2, NFκB, MAPK and IRF3. In addition, members of the NFκB complex (NFKBIE, NFKBIA, NFκB1 and RelA) as well as components of Nod1/2 and IFN signaling pathways (e.g. RIPK2, STAT1) were induced transiently at the vacuolar stage but back to normal levels at the cytosolic stage. Furthermore, genes involved in both pro-survival (Birc2, Birc3, Clarp2, CYR61) or apoptosis (TNFAIP8, PPID, ELMO2, HDAC1), ER and oxidative stress (ATF3, ATF4, ERCC2, XRCC5), cell cycle regulation (CDKN1A, Gadd45a, Clic-4), DNA damage repair (ERCC2, XRCC5, Gadd45a, HMGB1) and detoxification (CYP1B1) were induced. This gene expression pattern suggests the activation of a general stress response upon vacuolar presence of *Shigella*, possibly due to altered nutrient levels, autophagy and oxidative stress. Finally, genes involved in lipid metabolism (COX2, PTGER2) were also induced at the vacuolar stage. Thus, transcriptional activation of multiple PRR signaling pathways, of a general cellular stress response as well as of lipid metabolism hallmark the vacuolar signature of WT *Shigella*-infected cells.
4.3. Bulk cell results

**Cytosolic response**  As infection proceeded to the cytosolic stage, both sustained and differential regulation of early induced genes were observed. On the one hand, expression of genes from profile 1 and 2 remained elevated, namely genes involved in inflammation, stress, pro- and anti-apoptosis and detoxification. On the other hand, down-regulation of previously up-regulated genes was observed (profile 4), which was potentially triggered upon compartmental change of intracellular bacteria or was due to prolonged activity of bacterial effectors that dampen host gene transcription. Amongst others, this affected some, but not all, inflammation genes (NFKBIA, TNFAIP3, CXCL2, IL-8 and CCL20), which are part of the NFκB/MAPK-regulated proinflammatory signaling cascade. This suggested redundant regulation of proinflammatory genes through diverse signaling pathways, only some of which being affected by, respectively, the compartmental change or potential effector activity of the bacteria. Moreover, bacterial escape to the host cytosol led to de novo expression of additional genes. Expression of genes of profile 3, which included both pro- and anti-apoptosis genes (Caspase-4 and BECN1) as well as Cdc42, was induced. Also, the down-regulation of an additional set of genes was evident for the first time upon cytosolic bacterial localization (profile 5). This comprised genes involved in detoxification and DNA damage repair (CYP1B1, HMGB1), stress (Xbp1), inflammation (TLR4, STAT1, CTNNB1, NLRP1) autophagy and apoptosis (Atg16L, GPR43). Thus, the cytosolic signature of cells infected with WT *Shigella* is characterized by sustained expression of genes of inflammation, the stress response, repair, detoxification and apoptosis, but also by differential downregulation of a subset of inflammatory genes. Furthermore, this compartmental change induced de novo gene regulation, possibly contributing to a finely tuned adaptation of signaling pathways including the cellular stress response and detoxification as well as inflammation, apoptosis and autophagy.

4.3.4  The effector OspF has the strongest impact on immune responses

Next, we sought to investigate the impact of bacterial effectors on the intracellular transcriptional signatures. To investigate how bacterial effectors controlled localization dependent signatures, 'subtractive heatmaps' were generated. For each mutant, they display the relative differences of genes with a significant change (according to a Mann-Whitney-U test at $\alpha = 0.05$) between WT- and mutant-induced gene expression levels, thus shedding light on the impact of bacterial effectors on gene expression patterns. The heatmaps (see Figure 4.11) revealed that absence of OspF affected expression of more genes and more strongly so than any other absence of tested effectors. Significant differences of gene expression were most commonly observed for both vacuolar and cytosolic signatures, indicating that the impact OspF has on transcriptional signatures is largely independent of the bacterial subcellular localization. Moreover, although a few genes were less expressed in ΔospF-infected cells, indicating OspF-dependent induction of these genes, expression of the majority of investigated genes was significantly increased, indicating that the major role of OspF is to dampen host gene expression. OspF is known to inhibit MAPK signaling by irreversibly dephosphorylating p38 and ERK ([Arbibe et al., 2007]; [Li et al., 2007]). In line with this, targets of OspF-dependent inhibition comprised all genes of profile 4 and some of profile 1, i.e. proinflammatory chemokines (IL-8, CCL20, CXCL2) and other signaling components (NFKBIA, JunB, TNFAIP3, cJun) downstream of MAPK signaling,
4.3. Bulk cell results

Figure 4.11: Effector-dependency of responses to intracellular infection: The set of genes significantly differentially expressed between intracellular infection stages by WT and effector mutant *Shigella* strains are shown as heatmaps. Heatmap coloring indicates up- or down-regulation of the mutant response in comparison to WT *Shigella*. Significant changes at $\alpha = 0.05$ are indicated by + and − signs.

which were all significantly higher expressed in both vacuolar and cytosolic bacterial localization in the absence of OspF. Interestingly, OspF-dependent inhibition also included genes involved in the cellular integrated stress response, such as ATF3 and CDKN1A, as well as additional genes of apoptosis (Birc2), inflammatory signaling (RIPK1) and lipid metabolism (PTGER4), indicating an unexpected large-scale impact of OspF on host gene expression. Furthermore, expression of some genes of the stress response and repair (Xbp1, Rad17/51) as well as of the lipid metabolism (Akt-2) was decreased in the absence of OspF, indicating that OspF was required for gene regulation and this appeared to be dependent on the bacterial subcellular localization.

IpgD, is a PtdIns(4,5)P24-phosphatase, which amongst other effects leads to the release of extracellular ATP, thereby dampening host inflammation [Puhar et al., 2013] and to the activation of Akt kinase [Pendaries et al., 2006]. In conformity with these previous findings, we found strongly enhanced gene expression of a subset of inflammatory genes, including IFNs, CXCL10 and TLR4, as well as of a number of apoptosis genes (GPR43, CYR61) in the $\Delta ipgD$ mutant. Notably, these effects were restricted to the response upon cytosolic bacterial localization. In contrast, IpgD appeared to be required for induction of a subset of other inflammatory genes including NFKBIA, TNFAIP3 and IL-6 early upon intracellular infection and most likely independent of bacterial subcellular localization. Thus, IpgD mostly impacts the cytosolic intracellular signature and dampens expression of a subset of inflammatory and apoptosis genes.

The secondary effector OspG is known to inhibit downstream NFκB signaling by preventing degradation of IkB [Kim et al., 2005]. Our data revealed that proinflammatory gene expression was largely unaffected by infection with $\Delta ospG$, suggesting that proin-
flammatory signaling during *Shigella* infection was (redundantly) induced via pathways independent of NFκB. In contrast, expression of a number of genes related to apoptosis (Caspase-4, CYR61, HMGB1 and NFkB1), stress (Xbp1) as well as lipid metabolism (Akt-2) was significantly decreased in the absence of OspG, most notably in response to vacuolar bacterial localization. Taken together, OspG appears to be required for early induction of genes, which are largely involved in apoptosis, stress and lipid metabolism. MxiE is a transactivator of a secondary subset of effectors, including IpgD and OspG [Gall et al., 2005]. Our analysis of the MxiE mutant infection revealed significantly enhanced expression of a subset of inflammatory genes (IFN, CXCL10, HBD3) as well as genes involved in lipid metabolism (COX-2 and PTGER2) upon cytosolic bacterial localization, thus partly reflecting localization-dependent and -independent ∆ipgD-induced transcriptional responses. In contrast, a number of genes appeared to be MxiE-dependently expressed, e.g. genes of pro-apoptosis (NFκB2, PPID and ELMO2) and stress (Xbp1, CTNNB1), which partly reflects ∆ospG-induced gene expression patterns. Thus, the ∆mxiE-activated transcriptional response phenocopies a mixture of the localization dependently ∆ipgD-induced and the localization independently ∆ospG-induced signatures.

In summary, our analysis of bacterial effector mutants identified OspF as the major bacterial effector, which has a broad impact on intracellular transcriptional signatures largely independent of bacterial subcellular localization.

### 4.4 Single cell results

Applying the log-likelihood-ratio test defined in [Mcdavid et al., 2013] to the single cell data with very few exceptions reflected the bulk cell analysis results and thus further confirmed the comparability of single and bulk cell data. Most of the significant tests (>60%) were dominated by the proportional test component, while around 30% were dominated by the continuous component of the test and in about 10% of the tests both components contributed equally to the test result. The same trends were observed over all infection experiments with different *Shigella* strains. This indicates that most observed expression changes in the bulk cell data are mediated by the proportion of cells reacting to a given signal and not by a concomitant continuous increase of expression in all cells of a bulk cell population.

#### 4.4.1 Bimodal expression patterns

We thus decided to test whether single cell populations from individual infection stages showed bimodal behavior in terms of their expression. To this end, we applied a test strategy based on the *bimodality index* measure defined by [Wang et al., 2009]. Since the sample size for testing expression distribution shapes was very small (up to 40 cells measured, while many gene-infection stage pairs show significantly less measured samples), the only statistically significant results (α = 0.05) came from genes with two small, but clearly separated expression clusters (c.f. Figure A.7 in the appendix). While those few bimodally expressed genes might be interesting individually, the only systemic conclusion
4.4. Single cell results

we were able to draw from the bimodality tests, was that there were more genes showing bimodal behaviour in non-infected cells (13 genes) than in any of the other sample groups (7 to 8 genes) over all of the tested bacterial strains. Concordantly, the average variance of single cell expression values per gene was higher in the non-infected samples (1.4 ET^2), than in the other sample groups (1 ET^2). This increased variability of expression in non-infected samples indicates that they form a more heterogeneous population, potentially because they have been exposed to varying degrees of pathogen signals.

4.4.2 Single cell correlation analysis

To answer the question whether or not gene expression upon bacterial infection occurs in a coordinated manner, i.e. whether the expression of certain genes is jointly or independently regulated, we analyzed gene-pair correlations and assessed their statistical significance by merging evidence from the two individual experiments into a combined p-value.

Pathway-level results

The pairwise correlation p-values were used to assess the degree of correlation between genes assigned to a common biological pathway as detailed in section 4.2.2 as well as changes thereof between infection stages and between infections with different *Shigella* strains. Significantly increased (orange) or decreased (cyan) correlations within gene sets annotated to proinflammation, apoptosis and stress pathways were represented as heatmaps (Figure 4.12). This revealed that the degree of correlation among proinflammatory genes increased in bystander cells, but remained constant at a level similar to control in the course of WT intracellular infection (Figure 4.12A, top). In contrast, in the absence of OspF, correlations between proinflammatory genes significantly increased at both vacuolar and cytosolic stage of infection (Figure 4.12A, top and bottom). Furthermore, absence of OspG appeared to impact correlations at vacuolar bacterial localization, while absence of MxiE or IpgD showed no such impact. Dissection of individual gene-pair correlations identified a cluster of genes consisting of IL-8, CXCL-2, NFKBIA, IER3, CXCL-1 and cJun, whose correlations were largely lost at the cytosolic stage of WT and ΔospG, but not of ΔospF infection. This was further investigated for gene-pair correlations involving the proinflammatory chemokine IL-8 (Figure 4.12D). Pairwise correlations within this set of genes was significantly increased in bystander cells and at the vacuolar localization, followed by a significant decrease in correlation during the cytosolic stage of infection by almost all bacterial strains. However, upon infection with ΔospF, correlations remained significantly increased at the cytosolic stage, indicating that OspF impacts the coordinated expression of genes with IL-8.

Analysis of changes in the degree of correlation within apoptosis (Figure 4.12B) and stress gene pathways (Figure 4.12C) revealed that *Shigella* infection, in an OspF-dependent manner, led to both significantly increased correlation of apoptosis genes in cytosolic and of stress genes in vacuolar bacterial localization, thus implicating OspF not only in the deregulation of pro-inflammatory responses but also of apoptotic and stress responses.
4.4. Single cell results

Figure 4.12: Changes in degree of within-pathway correlation during infection: Represented are heatmaps displaying scores for the change in degree of correlation within the proinflammation (A), apoptosis (B) and stress (C) pathways as well as within a gene set centered around the pro-inflammatory cytokine IL-8 (D). Changes during the course of infection are displayed on the top row, the bottom row compares degree of correlation in distinct infection stages of effector-deficient mutants to the ones measured in WT *Shigella* infection. Significant correlation increases and decreases are displayed in orange and cyan, respectively.

**Correlation networks**

To get an unbiased view of correlation structures present in the data set, we visualized all gene-pairs detected as significantly (Benjamini-Hochberg multiple testing correction on $p_{comb}, \alpha = 0.01$) correlated as a separate network for each studied infection stage in Cytoscape. The full networks for each of the four sample groups of infection with WT *Shigella* and ∆ospF are shown in Figures 4.13 and 4.14.

They are characterized by a densely connected cluster of highly correlated genes which is stable over the course of infection and does not change with respect to different strains used for infection. This cluster is composed of highly expressed genes which seem to be correlated mainly due to cell-specific differences in total mRNA content. In control conditions, a second tightly correlated gene cluster, containing receptors (TLRs, GPR43), the interferon IFNa14 and the antimicrobial peptide HBD3, can be discerned. Genes in this cluster are expressed at a low level and only in very few — but the same — of the 40 measured single cells. This hints at a heterogeneous control sample population in which a few cells seem to be already primed for a defense reaction against potential pathogens. These cells also show decreased expression of some genes from the main cluster (ATF4, CLIC-4, IKKα, ITGB1 and IPS1 most notably) which manifests as anti-correlations in the network.

The TLR cluster is still detectable in non-infected cells of both WT and ∆ospF infection experiments. Induction of pathogen responses leads, however, to a third set of genes forming a new cluster, which — though not completely separate from the main cluster — shows high correlations between pro-inflammatory genes (IL-8, CXCL-2), NFKBIA,
4.4. Single cell results

NFkBIE, but also contains stress-response genes (ATF3, TNFAIP3). The measured correlations between the genes of this cluster are mediated by large differences in gene expression between individual single cells. Interestingly, under $\Delta ospF$ infection this infection response cluster contains more pro-inflammatory genes (IL-6, CXCL-1) and is more tightly connected than under the condition of WT infection. In the vacuolar infection stage, the genes of the infection response cluster are further induced, while correlations within the TLR cluster are lost. The infection response cluster is, like already observed in the non-infected stage, bigger and more tightly connected under $\Delta ospF$ than under conditions of WT infection. In the cytosolic infection stage, $\Delta ospF$ infection globally leads to fewer significant gene-pair correlations than WT infection. Pairwise correlations between the genes constituting the core of the infection response cluster (IL-8, CXCL-2, NFkBIA, TNFAIP3) are more pronounced upon cytosolic localization of $\Delta ospF$ bacteria as compared to WT infection.

Taken together, the correlation networks indicate heterogeneity within the measured single cell populations regarding the expression intensity of co-regulated subsets of genes. However, infection with the OspF-deficient mutant prompted strong, highly co-regulated pro-inflammatory responses, suggesting that this effector could be responsible for a disruption of regulatory pathways needed for a proper activation of immune responses.
4.4. Single cell results

Figure 4.13: Correlation networks during WT Shigella infection
4.4. Single cell results

Figure 4.14: Correlation networks during △ospF Shigella infection
4.4.3 Effect of infection on cell cycle phase distributions

The cell cycle phase predictions for replicate control samples were closely reproduced over individual arrays of an experiment. This indicates robustness of the predictions to technical fluctuations in the range of ±1 CT value. Furthermore, randomly permuting the expression values of the three cyclins independently for each single cell completely changed the proportions of predicted cell cycle phases in comparison to the characteristic proportions observed in control samples (see Figure A.8 in the appendix), indicating that e.g. the low percentage of S-phase cells (c.f. Figure 4.15) was characteristic for the control samples and not an artifact of the classification procedure. The comparison between cells derived from samples subjected to bacterial infection and the control samples showed that the overall distribution into cell cycle phases remained unchanged during each stage of infection with WT *Shigella*. In contrast, infection with the mutants ΔipgG, ΔospG and ΔospF showed significant changes in the proportion of predicted cell cycle phases. More specifically, cyclin A typically upregulated in the G2-phase showed higher expression in cells with cytosolic ΔipgG and ΔospG than in cells from control samples which lead to an increase in prediction of G2-phase in those cells. Non-infected cells taken from an infection experiment with ΔospF showed higher levels of cyclin B and were thus enriched for M-phase predictions (see Figure 4.15). The results presented above thus indicate the studied effectors to be directly or indirectly involved in the regulation of cell cycle control mechanisms.

4.5 Synopsis

We have established a workflow which combines fluorescence-based approaches compatible for FACS-based cell sorting and multiplex qPCR single cell analysis on a subset of host immune response genes using microfluidic devices. This enables us to discriminate and purify cells in distinct stages of bacterial infection with precise spatio-temporal resolution and to yield high-content, integrative information on transcriptional signatures of the host immune response of individual cells.

4.5.1 Summary

**Bulk cell analysis** Analysis of the transcriptomic data of small bulk-cell populations using a custom analysis pipeline revealed that distinct bacterial subcellular localization significantly correlates with differentially regulated transcriptional signatures of the host inflammatory and cellular stress response. We have found that immune responses of non-infected bystander cells are dependent on bacterial internalization in neighboring infected cells and are largely independent of the tested bacterial effectors. We classified gene responses into five distinct expression profiles and were thus able to show that genes previously believed to be regulated by the same mechanisms exhibit different expression profiles over the distinguished infection stages. Finally, using bacterial mutants known to impact specific transcriptional host responses, we identified the *Shigella* effector OspF to have a stronger impact on transcriptional immune response signatures than assumed and
Figure 4.15: Cell cycle phase changes during infection: (A) During infection with WT Shigella, no significant changes in the cell cycle phase predictions are observed. (B) Infection with ΔospF leads to more M-phase predictions in non-infected cells (C,D) Infection with ΔipgG and ΔospG leads to more G₂-phase predictions in cells with bacteria in the cytosol.
4.5. Synopsis

to be involved in the regulation of stress response and apoptosis genes in addition to its known role in dampening proinflammatory gene induction.

**Single cell analysis** Furthermore, we developed a set of statistical tests and data evaluation schemes specifically adapted to single cell analysis on small sample sizes and demonstrated that the single cell data, after adequate pre-processing, were highly concordant with bulk cell data. Evaluation of correlation structures in the single cell data revealed that immune responses were largely characterized by gene inductions in a subset but not in all of the measured single cells. In the population of non-infected cells, many genes showed a large degree of expression heterogeneity. Furthermore, we showed that OspF led to the disruption of an otherwise highly correlated module of pro-inflammatory genes induced in response to the infection. Prediction of cell cycle phases led to the hypothesis that the effectors OspF, IpgD and OspG might — directly or indirectly — be involved in the deregulation of the host cell cycle control.

4.5.2 Discussion

**Bulk cell analysis** As mentioned above, we observed a bystander response of non-infected cells that required internalization of bacteria in neighboring cells but showed no dependence of the tested bacterial effectors. Furthermore, this bystander response showed a strong overlap with the response to bacteria in vacuolar localization. Genes showing changes between the bystander and the vacuolar response were mostly characterized by a lower expression in the vacuolar localization. This indicates that the bystander effect might be a host immune strategy to circumvent immune suppression by bacterial pathogen effectors in infected cells by contributing to the amplification of responses on a population level. This strategy is concordant with the study by [Kasper et al., 2010] who observed that p38 activation and subsequent H3 phosphorylation in bystander cells were not affected by the effector OspF. As to the mode of bystander activation, Kasper et al. suggested that it was not mediated by paracrine signaling, such as TNF or other long-ranged diffusing soluble factors, but required cell-to-cell contact and involved gap junction signaling via the connexin Cx43 by studying Cx43-overexpressing A431 cells. Likewise, [Ablasser et al., 2013] showed, on HEK cell lines, that cGAMP spread to neighboring cells via gap junctions to promote the activation of STING in bystander cells. Here, we have shown that also *Shigella*-infected HeLa cells, known to be not connected through gap junctions ( [Eckert et al., 1993]; [Patel et al., 2009]), were clearly able to activate neighboring non-infected cells to produce a bystander immune response. Thus, alternative pathways of bystander activation independent of gap junctions have to be considered. Promising candidate molecules for mediating such signaling could be extracellular ATP [Puhar et al., 2013], Ca$^{2+}$ or inositol-3-phosphate.

Our study identified OspF as the major bacterial effector impacting transcriptional host immune responses. OspF is largely known for its irreversible impact on MAPK-dependent pro-inflammatory signaling and has been indicated particularly to act on the promoter of IL-8 ( [Arbibe et al., 2007]; [Li et al., 2007]). Here, we have shown OspF to have a much broader impact on host gene expression and highlight its global impact on gene expression of the stress response, including cell cycle and DNA repair, apoptosis and lipid
metabolism. Thus, the impact of OspF on global host gene expression has to be revisited. In this regard it is interesting to note that OspF has been shown to directly target the host chromatin leading to histone modifications and thereby modulating general host gene transcription ([Arbibe et al., 2007]; [Zurawski et al., 2009]). In contrast to ∆ospF infection, proinflammatory gene expression was much less affected upon infection with an OspG-deficient mutant. This suggests that proinflammatory signaling primarily involves MAPK signaling, which is a target of OspF-, but not of OspG-dependent inhibition.

**Single cell analysis** Despite the efforts taken to correct for biases due to small sample sizes, our test for bimodality still showed a lack of sensitivity leading to the prediction of very few bimodally expressed genes. Nevertheless, multiple observations, like the heterogeneity between individual cells in terms of the immune response induction and the fact that non-infected cells showed large differences in terms of the expression level of individual genes, point towards an inherent heterogeneity in the immune response of individual cells. Possible explanations for this could be the lack of synchronization of infections and that the measured populations — while showing identical subcellular localizations of bacteria — could still show variations in the timing of the infection. More intriguing explanations are intrinsic differences between single cells, e.g. in terms of their epigenetic state, their sensitivity in the perception of the bacterial stress (desensitization by previous stimuli, expression levels of receptors) or the cell cycle phase they are in.

The latter could be an interesting explanation as *Shigella* and other bacterial pathogens have been shown to directly control cell cycle progression through specific toxins and effector proteins ([Nougayrède et al., 2005]; [Iwai et al., 2007]). While the prediction of cell cycle phases indicated the involvement of *Shigella* effectors in host cell cycle regulation, it has to be noted that it depended on a very simple prediction tool taking into account only the relative levels of three cell cycle phase markers. The proportion of cell cycle phase assignments and changes thereof during infection have additionally been determined on small sample sizes (up to 40 individual single cells per infection stage) and can be expected to be subject to stochastic fluctuations. In order to test the specific effect of the indicated bacterial effectors on cell cycle regulation, further experimental work would thus be required. This could for example be achieved by a FACS gating strategy based on using propidium iodide (PI) staining in conjunction with the developed CCF4 assay. Such an approach would allow direct experimental quantification of the proportions of cells in a specific cell cycle phase and their dependence on the bacterial localization.

Generally, a more robust analysis of expression distributions over single cells would necessitate an increase in the number of measured single cell samples. Likewise, the measured gene-pair correlations, especially involving genes that are expressed only in a small subset of cells, could be made more robust by increasing the sample size. While for the pathway-level analysis, we developed tests to counteract these effects, they depend on predefined annotation classes and can thus only give a broad overview of correlation structures. Applying the Bayesian network learning tool Banjo [Hartemink et al., 2001] to the single cell data did not yield plausible regulatory networks (data not shown). This is not surprising due to the small sample size of 40 single cells, out of which a high proportion did not express many genes, and the larger number of variables (96 genes). Incorporating prior knowledge about e.g. non-transcriptional regulation or the mode of action of tested
4.5. Synopsis

effectors in conjunction with larger single cell samples, should lead to better inference performance.

In summary, we believe that the computational tools developed in this project have allowed the formulation of intriguing hypotheses and could — when applied to single cell data sets with larger sample sizes — lead to more detailed insights into the regulation mechanisms underlying the transcriptional host immune response to different stages of *Shigella* infection and how they are manipulated by *Shigella* effectors.


Bibliography


119


Bibliography


Bibliography
Appendix A

Appendix

A.1 Identification of additional proteins in differential proteomics

Detection of differentially expressed proteins  Bioinformatic analysis between 2D-DIGE images of 11 unaffected and 13 affected SMC protein extracts resulted in 569 polypeptide spots differentially profiled (p-value < 0.05). Of these, 408 polypeptide spots presented a fold change equal or higher than two. After manual validation that required a q-value < 0.05, 135 spots were selected as differentially profiled. Of these, 62 had increased and 73 had decreased abundance in 2D-DIGE gels of unaffected compared to affected SMC protein extracts. MALDI MS led to the identification of 78.5% of selected spots, i.e. 41 different human proteins from 100 spots, and seven spots corresponded to bovine DNase I, three of which were used for the classification of protein extracts.

Western blot analysis  Fifteen µL of 95% Laemmli buffer (2% SDS, 25% glycerol, 62.5 mM Tris HCl, 0.01% Bromophenol blue)/5% beta-mercaptoethanol were added to the volume corresponding to 50 µg of each SMC protein extract (ten unaffected and 11 affected), and incubated at 95,°C for 10 min. Denaturized samples were separated by 10% acrylamide SDS-PAGE and proteins were electrotransferred onto a 0.45-µm Hybond nitrocellulose membrane (GE Healthcare). Transferred proteins were incubated at 4,°C, overnight with primary antibodies, (monoclonal rat antihuman beta-arrestin 1 (1:150 v/v, R&D Systems, UK) and polyclonal goat antihuman beta-arrestin 2 (1:500 v/v, Abcam, UK)), that were diluted in 5% w/v nonfat dry milk in TBS-Tween. Incubation with secondary antibodies (donkey antigoat (Abcam) and ECL rabbit IgG-HRP (GE Healthcare)), diluted 1:5000 v/v in 5% w/v nonfat dry milk in TBS-Tween, was performed at room temperature for 1.5 h. Then, the specific proteins were detected using ECL Plus Western blotting detection reagent (GE Healthcare) followed by membrane scanning with an Ettan Dige Imager scanner (GE Healthcare) at excitation/emission wavelengths of 480 nm/530 nm to yield images with a pixel size of 100 µm. Finally, Quantity One software (Biorad, UK) was used for the acquisition of intensity values of detected proteins from blot images.
SUPPORTING INFORMATION

IDENTIFICATION OF ADDITIONAL PROTEINS IN DIFFERENTIAL PROTEOMICS USING PROTEIN INTERACTION NETWORKS

Frederik Gwinner⁴,⁵,⁶, Adelina E Acosta-Martín⁴,⁵,⁶,⁷, Ludovic Boytard⁴,⁵, Maggy Chwastyniak⁴,⁵,⁶, Olivia Beseme⁴,⁵,⁶, Hervé Drobecq⁶,⁷, Sophie Duban-Deweere⁷, Francis Juthier⁶, Brigitte Jude⁶, Philippe Amouyel⁴,⁵,⁶, Florence Pinet⁴,⁵,⁶,⁷, Benno Schwikowski⁴,⁵,⁶

¹INSERM, U744, Lille, France
²University of Lille Nord de France, USDL, Lille, France
³Institut Pasteur de Lille, Lille, France
⁴Systems Biology Laboratory, Institut Pasteur, Paris, France
⁵CNRS, URA2171, Paris, France
⁶CNRS, UMR8525, Lille, France
⁷E.A.2465, IMPRT-IFR114, University of Artois, Lens, France
⁸Centre Hospitalier régional et Universitaire de Lille, Lille, France
* Co-last authors
# These authors contributed equally to this work
¶Present address: Biomedical Proteomics Research Group, Faculty of Medicine, Geneva University, Geneva, Switzerland

CORRESPONDING AUTHOR: Benno Schwikowski, Systems Biology Laboratory, Institut Pasteur, 25-28 rue du Dr. Roux, 75015 Paris, France. Telephone: (+33) 1 45 68 86 20. Fax: (+33) 1 40 61 37 04. E-mail: benno@pasteur.fr
1. MATERIALS AND METHODS

1.1. Experimental background of the proteomic dataset

In a previous study in human monocytes derived macrophages obtained from donors over a four-year period [1], we had showed that the use of DNase I for the elimination of nucleic acids in proteomic cell samples produces unexpected proteome profile modifications. From technical and biological points of view, the mechanisms behind these proteome profile modifications are not understood, but they are highly relevant for laboratory practice since they potentially generate a bias in the analysis of protein expression patterns.

1.2. Isolation and culture of human smooth muscle cells

For a period of 5 years (2002-2006), 24 primary cultures of smooth muscle cells were prepared as previously described [2] adapted from Battle et al. [3]. Briefly, SMC were isolated from a residual segment of human internal mammary arteries obtained from 24 of the patients included in CORONA clinical protocol (CCPPRB n°CP 01/96 of 04/12/01) who underwent coronary artery bypass grafting in CHRU of Lille (France). The ethics committee of the Lille University Hospital Center (France) approved the study, and each patient provided written informed consent. The media was stripped from the underlying adventitia, finely minced and digested for 45 min at 37°C in 5 mL of HAM-F10 (Gibco BRL, Grand Island, New York, USA) containing 3 mg of collagenase (type I, 235 U/mg; Gibco BRL), 7 mg of elastase (3.73 U/mg; Worthington, Lakewood, NJ, USA) and 5 mg of soybean trypsin inhibitor (Sigma, St. Louis, MO, USA). Then, the enzymatic activity was stopped by adding 30% fetal calf serum (Bio-West, Miami, Florida, USA). Cultures of SMC were grown separately in HAM-F10 medium supplemented with 30% fetal calf serum.
serum, 10% human serum (Bio-West), 100 U/mL penicillin, 100 mg/mL streptomycin, 2.5 U/mL fungizone, 20 mM HEPES buffer (Gibco BRL), 2 mg/mL insulin (Roche Diagnostics, Mannheim, Germany), and 1% non essential amino acid solution (Gibco BRL). Medium was changed every 3 days. Cells were trypsinized at confluence and reseeded at a ratio 1:2. Confluent cells at passage 2 were washed three times with Hank’s Balanced Salt Solution buffer and then incubated for 24 h in serum-free culture medium before protein extraction was performed.

Human aortic SMC (ASMC) (Cat. No.: C-12532, PromoCell, Germany) were cultured in the same conditions as described above. Cells were trypsinized at confluence and reseeded at an estimated concentration of 0.4 million cells per dish. Proteins were extracted from confluence cells at passage 9. This ASMC sample is referred to as standard.

1.3. Extraction of intracellular proteins from human SMCs

SMCs were washed three times with 5 mL of 25 mM Tris, pH 7.4, and scraped into 75 µL of buffer containing 50 mM Tris pH 8.6, 10 mM EDTA, 65 mM DTT, proteinase-inhibitor cocktail (1 tablet for 10 mL buffer) (Complete, Roche Diagnostics, Meylan, France), 2,000 U/mL DNase I and 2.5 mg/mL RNase A (Roche Diagnostics) and incubated for 10 min at room temperature. Cells were then lysed in ice using a mixer suitable for 1.5 mL tubes and protein extracts were immediately stored at -20°C.

1.4. 2D-Difference Gel Electrophoresis

The 2D Clean-up kit (GE Healthcare) was used for DTT elimination in SMC protein extracts (50 µg). Protein pellets were then resuspended in 20 µL of lysis buffer (7 M urea, 2 M thiourea, 4% w/v CHAPS, 30 mM Tris, pH 8). The protein concentration was determined with the Bio-Rad RCDC Protein Assay (Biorad, UK) and BSA as the protein
standard. DTT-cleaned samples were reduced and labeled with DIGE saturation dyes [4] as previously described [1]. Briefly, the procedure for the analytical gels consisted in adjusting 5 µg of every cleaned sample to 9 µL with lysis buffer, reducing it by incubation with 1 µL of 2 mM Tris (2-carboxyethyl) phosphine at 37°C for 1 h, adding 2 µL of 2 mM cyanine (Cy3 or Cy5) and incubating for 30 min at 37°C. The reaction was stopped by the addition of 12 µL of buffer containing 7 M urea, 2 M thiourea, 4% w/v CHAPS, 130 mM DTT and 2% v/v Pharmalyte™, broad range pH 3-10 (GE Healthcare). The procedure for preparative gels consisted in adjusting 500 µg of every cleaned sample to 250 µL with lysis buffer, reducing it by incubation with 10 µL of 20 mM Tris (2-carboxyethyl) phosphine for 1 h at 37°C, adding 20 µL of 20 mM Cy3, and incubating for 30 min at 37°C. The reaction was stopped by adding 165.5 µL of the buffer used for the analytical gels.

Our DIGE labeling strategy included the use of the standard SMC sample in all 2D-DIGE gels to facilitate the alignment of corresponding images during bioinformatic analysis with SameSpots software. Before 2D gel electrophoresis, Cy3-labeled samples (SMC proteins from the 24 grafted patients) were mixed with the corresponding Cy5-labeled samples (24 replicates of the standard SMC culture proteins) and the total volume was adjusted to 450 µL with buffer containing 7 M urea, 2 M thiourea, 4% w/v CHAPS, 13 mM DTT and 1% v/v Pharmalyte™, broad range pH 3-10 (GE Healthcare). For the IEF step, IPG strip (24 cm, linear gradient pH 3-10) were rehydrated with 450 µL of labeled sample mixture in a Protean IEF cell system (Biorad, UK) for 24 h without applying any current. After rehydration, IEF was performed at 300 V for 3 h, and then at a gradient to 1000 V for 6 h, at a gradient to 8000 V for 3 h and finally at 8000 V for 3 h. After IEF, the IPG strips were incubated for 10 min at room temperature with equilibration buffer containing 6 M urea,
0.1 mM Tris-HCl pH 8, 30% v/v glycerol and 2% w/v SDS, then applied to the top of
12.5% isocratic Laemmli gels sealed with low melting temperature agarose (GE Healthcare). SDS-PAGE was run in an Ettan-Daltsix system (GE Healthcare) at 20°C and
at a constant voltage of 70 V overnight followed by 300 V until the bromophenol front
reached the bottom of the gel. All electrophoresis procedures (labeling, first and second
dimension) were performed in the dark.

1.5. Image acquisition and bioinformatic analysis

Gels cast between two low-fluorescence glass plates were scanned with an Ettan DIGE
Imager scanner (GE Healthcare) at excitation/emission wavelengths of 532 nm/580 nm for
Cy3 and 633 nm/670 nm for Cy5 to yield images with a pixel size of 100 µm.

Image analysis was performed with Progenesis SameSpots v3.0 software (Nonlinear
Dynamics, Newcastle upon Tyne, UK) equipped with a specific module for the analysis of
multiply stained gels. Images from the 24 SMC protein samples were combined with the
corresponding images of the standard sample into 24 virtual gels. Within each of these, the
image corresponding to the standard sample was selected as the reference image. Then,
after an overall reference image was chosen among the 24 images of the standard sample,
each virtual gel was aligned to the overall reference by aligning the images of the
respective standard samples. Spot volumes were then calculated and normalized in each
image. In the next step, the appropriate gels were selected for each group. Then, volume
differences in protein spots were analyzed. Spots were considered to have different
normalized spot volumes whenever the following three criteria were satisfied: 1) p-value <
0.05 (one-way ANOVA analysis), 2) fold change > 2, and 3) q-value < 0.05 (false
discovery rate).
1.6. In gel tryptic digestion and peptide extraction

Spots corresponding to differences in the proteome profiles of the two groups of samples were manually excised from a preparative gel. Then, digestion was performed as previously described [1]. Briefly, before digestion, spots were washed for 15 min with 100 µL of 50 mM ammonium bicarbonate solution and then twice for 15 min with 100 µL of 50 mM ammonium bicarbonate / 50% ACN. They were then dried after adding 100 µL of ACN for 10 min. After discarding the supernatant, tubes were left open for 10 min to complete solvent evaporation. They were then rehydrated with 12 µL of a solution containing 0.025% ProteasMAX™ Surfactant, Trypsin Enhancer (Promega) in 50 mM ammonium bicarbonate and 3 µL of 40 µg/mL Trypsin Gold (Promega) in 50 mM acetic acid. After overnight digestion at 37°C, peptide extraction was carried out in two steps according to the manufacturer’s protocol for ProteasMAX™ Surfactant, Trypsin Enhancer. Peptides were then purified, desalted with ZipTip C18 tips (Millipore Bedford, MA, USA) according to the manufacturer’s protocol, and eluted with 3 µL of 0.1% TFA/50% ACN.

1.7. MALDI MS and protein identification

For the acquisition of mass spectra from the extracted and desalted peptides, 0.5 µL of the peptide solution were mixed with 0.5 µL of matrix solution (5 mg/mL of CHCA dissolved in 0.1% TFA / 50% ACN) on the MALDI-TOF MS target. External calibration was performed with a peptide mixture resulting from the tryptic digest of BSA (0.5 µg/mL). MALDI-TOF MS was then performed with a Voyager DE STR mass spectrometer (PerSeptive Biosystems, Framingham, MA, USA) equipped with a 337.1 nm nitrogen laser and a delayed extraction facility (125 msec). All spectra were acquired in a positive ion reflectron mode under 20 kV voltage, 61% grid. Typically, 100 laser shots were recorded.
per sample. The mass spectra were then calibrated and the peak lists were generated using DataExplorer software (PerSeptive Biosystems, Framingham, MA, USA). Proteins were identified by PMF running the MASCOT web searcher (http://www.matrixscience.com/, Matrix Science, UK) against the NCBInr 20100312 (10570301 sequences; 3602205473 residues) with the following parameters: Taxonomy: Human (230372 sequences); Enzyme: Trypsin; Fixed modifications: CyDye-Cy3 (C); Variable modifications: Oxidation (M); Peptide Mass Tolerance: ± 50 ppm; Peptide Charge State: 1+; Max Missed Cleavages: 1.

When proteins could not be identified as described above, MALDI-TOF/TOF mass spectrometry was applied as follows: The MALDI target plate (AnchorChip™, Bruker Daltonics, Bremen, Germany) was covered with CHCA matrix (0.3 mg/ml in acetone:ethanol, 3:6 (v/v)). Peptide mixture was directly applied onto the CHCA matrix thin layer. The m/z measurements were performed in automatic mode using FlexControl™ 2.2 software on an Ultraflex™ TOF/TOF instrument (Bruker Daltonics, Bremen, Germany) in positive reflectron mode for MS acquisition or LIFT mode for MS/MS analysis. External calibration was performed using the peptide calibration standard kit (Bruker Daltonics, Bremen, Germany). Peak lists were generated from MS and MS/MS spectra using FlexAnalysis™ 2.4 software (Bruker Daltonics, Bremen, Germany). Database searches were performed via ProteinScape 1.3 (Bruker Daltonics, Bremen, Germany), through Mascot (Matrix Science Ltd, London, UK) using the following parameters: Database: SwissProt Release 57.3 (468851 sequences; 166149756 residues); Taxonomy: Human (20332 sequences); Enzyme: Trypsin; Fixed modifications: CyDye-Cy3 (C); Variable modifications: Oxidation (M); Peptide Mass Tolerance: ± 50 ppm; MS/MS fragment tolerance: ± 0.5 Da; Peptide Charge State: 1+; Max Missed Cleavages: 1.
If the protein search found more than one significant protein according to the MASCOT probability-based Mowse score, all possible results were considered. When peptides matched to multiple members of a protein family, identification corresponded to the protein identified with the highest number of unique peptides.

REFERENCES


### Supplementary Table 4. Comparison to the SteinerNet method

<table>
<thead>
<tr>
<th>Rank in result list</th>
<th>Rank in combined result list</th>
<th>Gene name</th>
<th>Terminal interactions</th>
<th>Score*</th>
<th>Contained in our result list</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steiner nodes found by SteinerNet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>GRB2</td>
<td>19</td>
<td>7.549</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>ARRB1</td>
<td>14</td>
<td>5.431</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>TRAF6</td>
<td>10</td>
<td>4.705</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>H2AFX</td>
<td>12</td>
<td>4.387</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>MAP3K7IP2</td>
<td>6</td>
<td>2.968</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>UBC</td>
<td>3</td>
<td>2.267</td>
<td>yes</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>TRAF2</td>
<td>6</td>
<td>2.245</td>
<td>no</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>CFL1</td>
<td>3</td>
<td>2.044</td>
<td>yes</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>PPP2R2B</td>
<td>7</td>
<td>1.878</td>
<td>no</td>
</tr>
<tr>
<td>10</td>
<td>13</td>
<td>RAF1</td>
<td>4</td>
<td>1.849</td>
<td>no</td>
</tr>
<tr>
<td>11</td>
<td>14</td>
<td>APP</td>
<td>4</td>
<td>1.818</td>
<td>yes</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>MYC</td>
<td>3</td>
<td>1.613</td>
<td>yes</td>
</tr>
<tr>
<td>13</td>
<td>16</td>
<td>THBS1</td>
<td>2</td>
<td>1.541</td>
<td>yes</td>
</tr>
<tr>
<td>14</td>
<td>17</td>
<td>EIF2AK3</td>
<td>2</td>
<td>1.27</td>
<td>yes</td>
</tr>
<tr>
<td>15</td>
<td>20</td>
<td>MCC</td>
<td>4</td>
<td>1.052</td>
<td>yes</td>
</tr>
<tr>
<td>16</td>
<td>21</td>
<td>PK2</td>
<td>3</td>
<td>1.048</td>
<td>no</td>
</tr>
<tr>
<td>17</td>
<td>22</td>
<td>HSPH1</td>
<td>1</td>
<td>0.999</td>
<td>no</td>
</tr>
<tr>
<td>18</td>
<td>23</td>
<td>COPS6</td>
<td>1</td>
<td>0.997</td>
<td>yes</td>
</tr>
<tr>
<td>19</td>
<td>24</td>
<td>TTN</td>
<td>2</td>
<td>0.862</td>
<td>no</td>
</tr>
<tr>
<td>20</td>
<td>25</td>
<td>DDX48</td>
<td>1</td>
<td>0.788</td>
<td>no</td>
</tr>
<tr>
<td>21</td>
<td>26</td>
<td>TSG101</td>
<td>1</td>
<td>0.757</td>
<td>yes</td>
</tr>
<tr>
<td>22</td>
<td>27</td>
<td>TP53</td>
<td>1</td>
<td>0.657</td>
<td>yes</td>
</tr>
<tr>
<td>23</td>
<td>28</td>
<td>CTNNB</td>
<td>1</td>
<td>0.609</td>
<td>no</td>
</tr>
<tr>
<td>24</td>
<td>30</td>
<td>PUF60</td>
<td>2</td>
<td>0.547</td>
<td>no</td>
</tr>
<tr>
<td>25</td>
<td>31</td>
<td>MED14</td>
<td>1</td>
<td>0.473</td>
<td>no</td>
</tr>
<tr>
<td>26</td>
<td>32</td>
<td>RPA2</td>
<td>1</td>
<td>0.43</td>
<td>no</td>
</tr>
<tr>
<td>27</td>
<td>33</td>
<td>PCNA</td>
<td>1</td>
<td>0.426</td>
<td>yes</td>
</tr>
<tr>
<td>28</td>
<td>34</td>
<td>CDC42</td>
<td>1</td>
<td>0.414</td>
<td>yes</td>
</tr>
<tr>
<td>29</td>
<td>35</td>
<td>CUL2</td>
<td>1</td>
<td>0.413</td>
<td>no</td>
</tr>
<tr>
<td>30</td>
<td>36</td>
<td>XRN2</td>
<td>1</td>
<td>0.378</td>
<td>yes</td>
</tr>
<tr>
<td>31</td>
<td>37</td>
<td>MED31</td>
<td>1</td>
<td>0.363</td>
<td>no</td>
</tr>
<tr>
<td>32</td>
<td>38</td>
<td>MED26</td>
<td>1</td>
<td>0.308</td>
<td>no</td>
</tr>
<tr>
<td>33</td>
<td>39</td>
<td>ASCC2</td>
<td>1</td>
<td>0.1</td>
<td>no</td>
</tr>
<tr>
<td>33</td>
<td>39</td>
<td>CHD3</td>
<td>1</td>
<td>0.1</td>
<td>no</td>
</tr>
<tr>
<td>33</td>
<td>39</td>
<td>EWSR1</td>
<td>1</td>
<td>0.1</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>SMURF1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>TCEB1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>ATM</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>SIP</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>MTA2</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>DNM1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>ITCH</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>SFPQ</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>FNBP1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>RHEB</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>p54</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>RNPS1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>DNM2</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>RHOC</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>USP13</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>KDM1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>BLM</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>RM1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>ARHGAP1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>PCBP1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>MDC1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>RNF11</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>PTBP1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>TMPO</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>MED29</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>MED8</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>MAGOH</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>NFRYB</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>RAP1B</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>SNRP116</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>SNRPB</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>PHB2</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>UBR5</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>SNRPE</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>LSM3</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>MYBBP1A</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>UBXN1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>MCRS1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>42</td>
<td>NONO</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
</tbody>
</table>

Steiner nodes only found by our method

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>ARRB2</td>
<td>14</td>
<td>6.130613322</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>EGFR</td>
<td>8</td>
<td>3.755661566</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>VHL</td>
<td>5</td>
<td>1.957528849</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>18</td>
<td>MYOZ1</td>
<td>2</td>
<td>1.254514885</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>19</td>
<td>TSC1</td>
<td>3</td>
<td>1.227937152</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>29</td>
<td>XRN1</td>
<td>1</td>
<td>0.582513398</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>42</td>
<td>EXOSC10</td>
<td>0</td>
<td>0</td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>

* : The ranking score is computed as the sum of rfunsimBP scores of all edges connecting a given protein to any terminal node protein
A.1. Identification of additional proteins in differential proteomics

Figure A.1: Supplementary figure 4 from [Gwinner et al, Proteomics 2013], Significance analysis of the computed Steiner tree: 10,000 simulations were performed using 41 randomly selected proteins as input to the Steiner tree algorithm. A: Frequency distribution of the number of found Steiner nodes. B: Frequency distribution of the sum of Steiner tree edge costs. Gauss curves, in green, show that a normal distribution of the results can be assumed. Blue lines indicate the 95% confidence interval of the normal distributions. Red lines indicate values corresponding to the Steiner tree solution obtained when using the proteins detected in the 2D-DIGE approach as input, for which the number of Steiner nodes (p-value = 3.6 * 10^{-5}) and the sum of edge costs (p-value = 2.3 * 10^{-4}) are far lower than expected from the distributions of the simulation results.
A.1. Identification of additional proteins in differential proteomics

Figure A.2: Supplementary figure 5 from [Gwinner et al, Proteomics 2013], Distribution of biochemical properties for proteins identified by 2D-DIGE analysis and beta-arrestins: Tabulated values for the percentage of cysteines, number of cysteines, molecular weight (MW), isoelectric point (pI) and hydrophobicity (GRAVY) for the 41 identified human proteins (blue dots) and for the two beta-arrestins (larger yellow dots)
A.2 Plant responses to environmental stresses

Supplementary figures

Figure A.3: Evaluation of number of clusters on the root data set:
A.2. Plant responses to environmental stresses

Figure A.4: Evaluation of number of clusters on the shoot data set:

Figure A.5: Inferred regulatory network for the shoot tissue data. Clusters are represented as nodes, with node size indicating the number of contained genes. Strong regulatory interactions (absolute coefficients of at least 0.2) are depicted by edges (red: down-regulation; green up-regulation). Edge width indicates the strength of inferred regulation.
A.2. Plant responses to environmental stresses

Supplementary code

```matlab
%% weighted fitting test case
n_tests = 200;
% draw data from a preset impulse model
x = [0, 0.25, 0.5, 1, 3, 4, 6, 12, 24];
T = numel(x)
p_star = [0, 1, 0.5, 0.3, 3.5, 3];
y_star = impulse(p_star,x);
fitdiffs = zeros(2,n_tests);
for n=1:n_tests
    % add noise to create replicate measurements
    y_rep = [y_star ; y_star];
    noise_sd = rand( [1, T] ) * 0.2;
    for t = 1 : T
        y_rep(:, t) = y_rep(:, t) + randn( [2, 1] ) * noise_sd(t);
    end
    % compute mean of replicates
    y_mod = mean( y_rep );
    % compute weights
    weights = weights_from_range( y_rep );
    % run parameter estimations with and without weights
    [wparams, werr] = best_impulse_params( x,y_mod,100,3,weights );
    [params , err] = best_impulse_params( x,y_mod,100,1 );
    fitdiffs(1,n) = sum( (impulse(wparams,x) - y_star).^2 );
    fitdiffs(2,n) = sum( (impulse(params,x) - y_star).^2 );
end
% test whether weighted fitting improves model fitness
% using a one-tailed paired t-test at alpha = 0.05
[h,p,ci] = ttest( fitdiffs(1,:),fitdiffs(2,:),0.05,'left' )
```

Code A.1: Test case for weighted versus standard impulse model fitting
A.3 Single-cell transcriptomic signatures elicited by Shigella infection

Experimental protocols

**Bacterial strains and bacterial preparation**  *S. flexneri* M90T wild type (WT) and mutant strains deficient for the bacterial effectors OspF (ΔospF), OspG (ΔospG), MxiE (ΔmxiE), IpgD (ΔipgD) and MxiD (ΔmxiD) were transformed by electroporation with the pGG2(AmpR)-dsRed plasmid. For infection experiments, bacteria from frozen stocks were grown over night (ON) in tryptic casein soy broth (TCSB) supplemented with ampicillin (50 µg/ml) at 37°C with shaking at 220 rpm and subsequently sub-cultured at a 1/100 dilution for 2.25 h to an optical density at 600 nm (OD600) of > 0.3.

**Cell culture and transfection procedures**  All cell culture reagents were purchased from Invitrogen unless stated otherwise. Human epithelial HeLa cells (ATCC) were grown in Dulbecco’s modified Eagle’s medium (DMEM) containing 10% decomplemented fetal calf serum (FCS) at 37°C, 5% CO2. Transfection of HeLa cells with p-mOrange-empty or p-mOrange-bla was performed 48 h prior to experimentation using Fugene transfection reagent (Promega) according to the manufacturer’s instructions.

**FRET assay and infection of HeLa cells**  All infection experiments using the FRET assay were carried out in EM medium (120 mM NaCl, 7 mM KCl, 1.8 mM CaCl2, 0.8 mM MgCl2, 5 mM glucose and 25 mM HEPES at pH 7.3) and all buffers and media used until read-out were supplemented with 2.5 µM probenicid. For the FRET assay and subsequent infection, 2 × 10^5 HeLa cells were seeded per 12-well 24 h prior to infection. Subsequently, cells were washed 3x with PBS and loaded with 0.05 µM CCF4/AM substrate (Invitrogen) in EM containing 2.5 µM probenicid and 20% (v/v) Solution B for 2 h at room temperature (RT) in the dark. Bacteria were washed with PBS, coated for 10 min at RT with 10 µg/ml poly-L-lysine (Sigma) and 10 µg/ml soluble β-lactamase to facilitate bacterial adhesion to cells and enhance their enzymatic activity, respectively, and resuspended in EM medium. For subsequent infection, cells were washed 3x with PBS and incubated with EM containing indicated bacteria. Infection of HeLa cells was carried out at a multiplicity of infection (MOI) of 25 for 15 min at RT to allow bacterial attachment and synchronization of infection, followed by incubation for 30 min at 37°C to promote bacterial invasion. Subsequently, cells were extensively washed and further incubated for 2.5 h at 37°C, giving a 50-70% infection rate. Infection was stopped by maintaining cells at 4°C until read-out, which also stopped further transcriptional activity, since even several hours of cells on ice had no impact on overall gene expression signatures. For cell sorting, cells were trypsinized for 5 min at 37°C, centrifuged for 5 min at 200 * g and 4°C, resuspended in cold EM buffer and kept on ice until sort. For Imagestream analysis, trypsinized cells were fixed in 4% PFA for 30 min and then resuspended in PBS.

**FACS-based cell sorting, sample processing and multiplex qPCR**  Single cells or pools of 20 cells of desired populations were sorted on a FACS Aria III cell sorter (BD
A.3. Single-cell transcriptomic signatures elicited by *Shigella* infection

Bioscience) into 96-well plates containing 9 µl lysis/RT-STA mix for subsequent reverse transcription (RT) and specific target pre-amplification (STA) (per well: 5 µl Cells direct 2x Reaction Mix and 0.2 µl SuperScript TM III RT PlatinumR Taq Mix [CellsDirect One-Step qRT-PCR Kit, Life Technologies], 2 units Superase-In RNase Inhibitor [Life Technologies], 2.5 µl assay mix of 96 pooled Solaris qPCR gene expression assays [ThermoFisher] at a final concentration of 0.2x and 1.2 µl 0.1 mM EDTA/10 mM Tris buffer). After sorting, samples were vortexed, briefly centrifuged and subjected to one-step RT-STA using the following program: 50°C for 15 min, 95°C for 2 min (for RT) and 21-23 cycles of 95°C for 15 sec and 60°C for 4 min (for STA). For multiplex qPCR on a BioMark System (Fluidigm), cDNA was diluted 1/5 with 1 mM EDTA/10 mM Tris buffer and 2.9 µl cDNA was mixed with 3.25 µl 2x Solaris Universal qPCR MasterMix (ThermoFisher) and 0.32 µl Sample Loading Reagent (Fluidigm) and loaded together with 48 or 96 individual gene expression assays on a 48.48 or 96.96 Dynamic Array, respectively, according to the manufacturer’s instructions. Ct values were obtained using the BioMark Real-time PCR Analysis software (Fluidigm) and further used for subsequent analysis.

Supplementary tables

The pathways tested for significant changes in terms of overall degree of within-pathway correlations were composed as follows:

- **Pro-inflammatory genes**
  - NFkB1, c-Jun, NFkB2, NFKBIA, NFKBIE, RIPK1, IPS1, RIPK2, IKKa, NLRP1, Nod1, MDA5, TLR2, TLR4, TLR3, TLR6, IL-8, TNFa, IFNa14, IFNb, IL-6, HBD3, CCL-2, CCL20, CXCL-1, CXCL-2, IL-1a, IL-18, CXCL10, COX-2, RegIIIbeta, HMGB1, Caspase-1, Caspase-4, Caspase-5, TNFAIP3, STAT1, SOCS-1, UBP43, OAS1, IER3, HDAC1

- **Apoptosis genes**
  - NFkB1, c-Jun, NFKBIA, CHOP, RIPK1, NLRP1, Nod1, TNFa, IFNb, IL-1a, HMGB1, Caspase-1, Caspase-4, Caspase-5, TNFAIP3, STAT1, PPID, ALDH1A3, ERCC2, CDKN1A, Gadd45a, HDAC1, SMAD3, p53, ELMO2, ELMO1, XRCC5, IER3, Akt-2, Bnip3, RELA, GPR43, TNFAIP8, BECN1, Bcl2, CLARP2, Birc2, Birc3, CYR61, Fut-2

- **Stress response genes**
  - JUNB, c-Jun, NFKBIA, NFKBIE, CHOP, ATF3, ATF4, Xbp1, RIPK1, IL-8, IFNa14, IFNb, IL-6, CCL-2, CCL20, CXCL-1, CXCL-2, IL-1a, IL-18, HMGB1, Caspase-1, Caspase-4, Caspase-5, TNFAIP3, CALR, CANX, ERCC2, RAD17/51, CRY1, ALDH1A3, CYP1A1, CYP1B1, SOD1, CDKN1A, PPP2R1B, CCNE1, CCNB1, CCNA2, atg16L1, atg2A, XRCC5, BECN1, Bcl2
# A.3. Single-cell transcriptomic signatures elicited by *Shigella* infection

<table>
<thead>
<tr>
<th>Accession number</th>
<th>Gene name</th>
<th>Major annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM-000291</td>
<td>PGK1</td>
<td>reference gene</td>
</tr>
<tr>
<td>NM-000994</td>
<td>B2M</td>
<td>reference gene</td>
</tr>
<tr>
<td>NM-000034</td>
<td>SDHA</td>
<td>reference gene</td>
</tr>
<tr>
<td>NM-003998</td>
<td>NFkB1</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>NM-002229</td>
<td>JUNB</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>NM-002228</td>
<td>c-Jun</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>NM-002502</td>
<td>NFkB2</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>NM-020529</td>
<td>NFKBIA</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>NM-004556</td>
<td>NFKBIE</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>NM-004083</td>
<td>CHOP</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>NM-001674</td>
<td>ATF3</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>NM-182810</td>
<td>ATF4</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>NM-005080</td>
<td>Xbp1</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>NM-003804</td>
<td>RIPK1</td>
<td>PRR signaling</td>
</tr>
<tr>
<td>NM-020746</td>
<td>IPS1</td>
<td>PRR signaling</td>
</tr>
<tr>
<td>NM-003821</td>
<td>RIPK2</td>
<td>PRR signaling</td>
</tr>
<tr>
<td>NM-001278</td>
<td>IKKa</td>
<td>PRR signaling</td>
</tr>
<tr>
<td>NM-00103053</td>
<td>NLRP1</td>
<td>PRR and signaling</td>
</tr>
<tr>
<td>NM-006092</td>
<td>Nod1</td>
<td>PRR</td>
</tr>
<tr>
<td>NM-022168</td>
<td>MDA5</td>
<td>PRR</td>
</tr>
<tr>
<td>NM-003264</td>
<td>TLR2</td>
<td>PRR</td>
</tr>
<tr>
<td>NM-138554</td>
<td>TLR4</td>
<td>PRR</td>
</tr>
<tr>
<td>NM-003265</td>
<td>TLR3</td>
<td>PRR</td>
</tr>
<tr>
<td>NM-006068</td>
<td>TLR6</td>
<td>PRR</td>
</tr>
<tr>
<td>NM-005584</td>
<td>IL-8</td>
<td>pro-inflammation</td>
</tr>
<tr>
<td>NM-005594</td>
<td>TNFa</td>
<td>pro-inflammation</td>
</tr>
<tr>
<td>NM-002172</td>
<td>IFNa14</td>
<td>pro-inflammation</td>
</tr>
<tr>
<td>NM-002176</td>
<td>IFNb</td>
<td>pro-inflammation</td>
</tr>
<tr>
<td>NM-000600</td>
<td>IL-6</td>
<td>pro-inflammation</td>
</tr>
<tr>
<td>NM-018661</td>
<td>HBD3</td>
<td>pro-inflammation</td>
</tr>
<tr>
<td>NM-002982</td>
<td>CCL2</td>
<td>pro-inflammation</td>
</tr>
<tr>
<td>NM-00130046</td>
<td>CCL20</td>
<td>pro-inflammation</td>
</tr>
<tr>
<td>NM-001511</td>
<td>CXCL-1</td>
<td>pro-inflammation</td>
</tr>
<tr>
<td>NM-002089</td>
<td>CXCL-2</td>
<td>pro-inflammation</td>
</tr>
<tr>
<td>NM-000575</td>
<td>IL-1a</td>
<td>pro-inflammation</td>
</tr>
<tr>
<td>NM-001562</td>
<td>IL-18</td>
<td>pro-inflammation</td>
</tr>
<tr>
<td>NM-001565</td>
<td>CXCL10</td>
<td>pro-inflammation</td>
</tr>
<tr>
<td>NM-000963</td>
<td>COX-2</td>
<td>inflammation</td>
</tr>
<tr>
<td>NM-138937</td>
<td>RegIIIbeta</td>
<td>inflammation</td>
</tr>
<tr>
<td>NM-002128</td>
<td>HMGB1</td>
<td>inflammation</td>
</tr>
<tr>
<td>NM-0033292</td>
<td>Caspase-1</td>
<td>inflammation</td>
</tr>
<tr>
<td>NM-001225</td>
<td>Caspase-4</td>
<td>inflammation</td>
</tr>
<tr>
<td>NM-004347</td>
<td>Caspase-5</td>
<td>inflammation</td>
</tr>
<tr>
<td>NM-006290</td>
<td>TNFAIP3</td>
<td>anti-inflammation</td>
</tr>
<tr>
<td>NM-139266</td>
<td>STAT1</td>
<td>IFN signaling</td>
</tr>
</tbody>
</table>
### A.3. Single-cell transcriptomic signatures elicited by *Shigella* infection

<table>
<thead>
<tr>
<th>Gene Accession</th>
<th>Gene Symbol</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM-003745</td>
<td>SOCS-1</td>
<td>IFN signaling</td>
</tr>
<tr>
<td>NM-017414</td>
<td>UBP43</td>
<td>IFN signaling</td>
</tr>
<tr>
<td>NM-002534</td>
<td>OAS1</td>
<td>IFN signaling</td>
</tr>
<tr>
<td>NM-005038</td>
<td>PPID</td>
<td>heat shock resp.</td>
</tr>
<tr>
<td>NM-004343</td>
<td>CALR</td>
<td>cytoskeleton</td>
</tr>
<tr>
<td>NM-001746</td>
<td>CANX</td>
<td>cytoskeleton</td>
</tr>
<tr>
<td>NM-000400</td>
<td>ERCC2</td>
<td>DNA repair</td>
</tr>
<tr>
<td>NM-002873</td>
<td>RAD17/51</td>
<td>DNA repair</td>
</tr>
<tr>
<td>NM-004075</td>
<td>CRY1</td>
<td>DNA repair</td>
</tr>
<tr>
<td>NM-000693</td>
<td>ALDH1A3</td>
<td>detoxification</td>
</tr>
<tr>
<td>NM-00499</td>
<td>CYP1A1</td>
<td>detoxification</td>
</tr>
<tr>
<td>NM-00104</td>
<td>CYP1B1</td>
<td>detoxification</td>
</tr>
<tr>
<td>NM-00454</td>
<td>SOD1</td>
<td>detoxification</td>
</tr>
<tr>
<td>NM-013943</td>
<td>CLIC-4</td>
<td>cytoskeleton</td>
</tr>
<tr>
<td>NM-001904</td>
<td>CTNNB1</td>
<td>cytoskeleton</td>
</tr>
<tr>
<td>NM-001791</td>
<td>Cdc42</td>
<td>cytoskeleton</td>
</tr>
<tr>
<td>NM-078467</td>
<td>CDKN1A</td>
<td>cell cycle</td>
</tr>
<tr>
<td>NM-181699</td>
<td>PPP2R1B</td>
<td>cell cycle</td>
</tr>
<tr>
<td>NM-001238</td>
<td>CCNE1</td>
<td>cell cycle</td>
</tr>
<tr>
<td>NM-031966</td>
<td>CCNB1</td>
<td>cell cycle</td>
</tr>
<tr>
<td>NM-001237</td>
<td>CCNA2</td>
<td>cell cycle</td>
</tr>
<tr>
<td>NM-017974</td>
<td>Atg16L1</td>
<td>autophagy</td>
</tr>
<tr>
<td>NM-015104</td>
<td>Atg2A</td>
<td>autophagy</td>
</tr>
<tr>
<td>NM-004964</td>
<td>HDAC1</td>
<td>apoptosis</td>
</tr>
<tr>
<td>NM-001924</td>
<td>Gadd45a</td>
<td>apoptosis</td>
</tr>
<tr>
<td>NM-005902</td>
<td>SMAD3</td>
<td>apoptosis</td>
</tr>
<tr>
<td>NM-000546</td>
<td>p53</td>
<td>apoptosis</td>
</tr>
<tr>
<td>NM-133171</td>
<td>ELMO2</td>
<td>apoptosis</td>
</tr>
<tr>
<td>NM-014800</td>
<td>ELMO1</td>
<td>apoptosis</td>
</tr>
<tr>
<td>NM-021141</td>
<td>XRCC5</td>
<td>apoptosis</td>
</tr>
<tr>
<td>NM-003897</td>
<td>IER3</td>
<td>apoptosis</td>
</tr>
<tr>
<td>NM-001626</td>
<td>Akt-2</td>
<td>apoptosis</td>
</tr>
<tr>
<td>NM-004052</td>
<td>Bnip3</td>
<td>apoptosis</td>
</tr>
<tr>
<td>NM-021975</td>
<td>RELA</td>
<td>apoptosis</td>
</tr>
<tr>
<td>NM-005306</td>
<td>GPR43</td>
<td>apoptosis</td>
</tr>
<tr>
<td>NM-014350</td>
<td>TNFAIP8</td>
<td>apoptosis</td>
</tr>
<tr>
<td>NM-003766</td>
<td>BECN1</td>
<td>anti-apoptosis</td>
</tr>
<tr>
<td>NM-000633</td>
<td>Bcl2</td>
<td>anti-apoptosis</td>
</tr>
<tr>
<td>NM-003879</td>
<td>CLARP2</td>
<td>anti-apoptosis</td>
</tr>
<tr>
<td>NM-001166</td>
<td>Birc2</td>
<td>anti-apoptosis</td>
</tr>
<tr>
<td>NM-001165</td>
<td>Birc3</td>
<td>anti-apoptosis</td>
</tr>
<tr>
<td>NM-001554</td>
<td>CYR61</td>
<td>adhesion</td>
</tr>
<tr>
<td>NM-033668</td>
<td>ITGB1</td>
<td>adhesion</td>
</tr>
<tr>
<td>NM-002205</td>
<td>a5-Integrin</td>
<td>adhesion</td>
</tr>
<tr>
<td>NM-000201</td>
<td>ICAM-1</td>
<td>adhesion</td>
</tr>
<tr>
<td>NM-000956</td>
<td>PTGER2</td>
<td>lipid metabolism</td>
</tr>
</tbody>
</table>
### A.3. Single-cell transcriptomic signatures elicited by *Shigella* infection

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene Symbol</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM-000958</td>
<td>PTGER4</td>
<td>lipid metabolism</td>
</tr>
<tr>
<td>NM-004878</td>
<td>PTGES</td>
<td>lipid metabolism</td>
</tr>
<tr>
<td>NM-002332</td>
<td>LRP-1</td>
<td>lipid metabolism</td>
</tr>
<tr>
<td>NM-000511</td>
<td>Fut-2</td>
<td>metabolism</td>
</tr>
<tr>
<td>NM-006202</td>
<td>Pde4A</td>
<td>membrane ruffle</td>
</tr>
<tr>
<td>NM-004815</td>
<td>ARHGAP29</td>
<td>GTPase-act. protein</td>
</tr>
<tr>
<td>NM-138551</td>
<td>TSLP</td>
<td>commensal interaction</td>
</tr>
<tr>
<td>NM-024111</td>
<td>CHAC1</td>
<td>stress response</td>
</tr>
<tr>
<td>NM-004836</td>
<td>PERK</td>
<td>stress response</td>
</tr>
<tr>
<td>NM-1013703</td>
<td>GCN2</td>
<td>stress response</td>
</tr>
</tbody>
</table>

Table A.1: Table of genes measured on the Biomark qPCR arrays

---

**Supplementary figures**

**Figure A.6:** Separation between sample groups in PCA: (A) PCA plot computed over 20 genes significantly differentially expressed between control and non-infected samples. (B) PCA plot computed over 15 genes significantly differentially expressed between vacuolar and cytosolic infection samples.
A.3. Single-cell transcriptomic signatures elicited by *Shigella* infection

Figure A.7: Genes detected as bimodally expressed: Shown are violin plots of genes with expression detected as significantly deviating from unimodal distribution assumption during infection with wild-type *Shigella*. Significances at $\alpha = 0.05$ (*), 0.01 (**) and 0.001 (***) according to the adapted bimodality index test described in section 4.2.2 are indicated on the x-axis.
A.3. Single-cell transcriptomic signatures elicited by *Shigella* infection

Figure A.8: Cell cycle phases predicted on randomly permuted data: The characteristic proportions between predicted cell cycle phases in control samples shown in Figure 4.15 are lost when the expression values of the three cyclins are randomly permuted independently for each cell.