Doctoral Thesis

Statistical inference on pathogen entry into human cells

Author(s): 
Drewek, Anna M.

Publication Date: 
2016

Permanent Link: 
https://doi.org/10.3929/ethz-a-010607487

Rights / License: 
In Copyright - Non-Commercial Use Permitted
STATISTICAL INFERENCE ON PATHOGEN ENTRY INTO HUMAN CELLS

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
ANNA MICHALENA DREWEK

Master of Sciences, ETH Zurich
born on 21.03.1986
citizen of Germany

accepted on the recommendation of

Prof. Dr. Peter Bühlmann, examiner
Prof. Dr. Niko Beerenwinkel, co-examiner

2016
Contents

Abstract vii
Zusammenfassung ix
Abbreviations xi

1 Introduction 1
  1.1 Infectious Diseases ........................................ 1
  1.2 Outline .................................................. 3

2 Biological and Experimental Background 7
  2.1 Molecular Biology .......................................... 7
  2.2 Inference with SiRNAs .................................... 9
  2.3 Pathogen Entry ............................................ 10
    2.3.1 Viral Entry ........................................ 10
    2.3.2 Bacterial Entry ...................................... 11
    2.3.3 Studied Pathogens .................................... 11
  2.4 InfectX Experiments ...................................... 13
    2.4.1 RNAi Screens ........................................ 13
    2.4.2 From Images to Data .................................. 15
    2.4.3 Infection Scoring .................................... 17

Appendix
  2.A Microscopy .............................................. 18
  2.B Computational Infrastructure ............................ 18
  2.C Image Analysis .......................................... 19
    2.C.1 Object Detection and Feature Extraction ..... 19
    2.C.2 Methods for Infection Scoring ...................... 20


2.C.3 Infection Scoring for Each Pathogen  

3 Simultaneous Analysis of large-scale RNAi Screens  

<table>
<thead>
<tr>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Background</td>
</tr>
<tr>
<td>3.2 Results</td>
</tr>
<tr>
<td>3.2.1 Kinome-Wide RNAi Screening for Infection</td>
</tr>
<tr>
<td>3.2.2 Data Reproducibility</td>
</tr>
<tr>
<td>3.2.3 Parallel Mixed Model (PMM)</td>
</tr>
<tr>
<td>3.2.4 Analysis of SiRNA Libraries</td>
</tr>
<tr>
<td>3.2.5 Sharedness of Detected Significant Genes</td>
</tr>
<tr>
<td>3.2.6 Comparison to Existing Ranking Methods</td>
</tr>
<tr>
<td>3.3 Additional Analysis Results</td>
</tr>
<tr>
<td>3.3.1 Cell Count Readout</td>
</tr>
<tr>
<td>3.3.2 Similarities and Differences</td>
</tr>
<tr>
<td>3.3.3 False Negatives</td>
</tr>
<tr>
<td>3.3.4 Parallelism of Screens</td>
</tr>
<tr>
<td>3.3.5 Phylogenetic Relations</td>
</tr>
<tr>
<td>3.4 Methods</td>
</tr>
<tr>
<td>3.4.1 Data Preprocessing and Normalization</td>
</tr>
<tr>
<td>3.4.2 Access to Data</td>
</tr>
<tr>
<td>3.4.3 Parallel Mixed Model (PMM)</td>
</tr>
<tr>
<td>3.4.4 Moderated T-Test</td>
</tr>
<tr>
<td>3.4.5 Redundant SiRNA Analysis</td>
</tr>
<tr>
<td>3.4.6 Power Analysis of PMM</td>
</tr>
<tr>
<td>3.4.7 Test for Stability</td>
</tr>
<tr>
<td>3.4.8 Hit Reproducibility</td>
</tr>
<tr>
<td>3.4.9 Goodness of Gene Rankings</td>
</tr>
<tr>
<td>3.4.10 Influence of Parallelism</td>
</tr>
<tr>
<td>3.4.11 Gene Set Enrichment Analysis</td>
</tr>
<tr>
<td>3.5 Biological Inquiry on Detected Significant Genes</td>
</tr>
<tr>
<td>3.6 Discussion and Conclusions</td>
</tr>
</tbody>
</table>

Appendix

<table>
<thead>
<tr>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.A Biological Protocols</td>
</tr>
<tr>
<td>3.A.1 Assays Performed</td>
</tr>
<tr>
<td>3.A.2 Main Differences in Protocols</td>
</tr>
<tr>
<td>3.A.3 Wet-lab Protocols</td>
</tr>
</tbody>
</table>
6 Conclusions and Outlook
   6.1 Theoretical Findings .................................. 119
   6.2 Applicational Findings .............................. 121

Bibliography ............................................. 123
Abstract

Infectious diseases caused by bacteria or viruses are among the top 10 global causes of death. Although in the last century many medical treatments have successfully been developed, bacteria and viruses are still of major public concern since an increasing number of them has acquired resistant mechanisms against the existing drugs. New approaches are required to prevent these resistant pathogens from entering human cells.

A comprehensive understanding of the entry process is substantial for the development of new mechanisms against pathogens. Large-scale RNAi screenings provide a powerful tool for an exhaustive investigation of biological pathways on a wide range of cellular processes, including the entry pathway of pathogens. This relatively new technique generates in most cases very noisy, high-dimensional data. Exploiting these large amount of data is challenging. In this thesis we propose several approaches for hit detection and cross-comparison of RNAi screens, as well as one approach for the reconstruction of the partial correlations between the interfered genes.

As a first approach we developed the Parallel Mixed Model (PMM) which is suitable for hit selection and cross-comparison of RNAi screens performed in parallel under several conditions. PMM simultaneously analyzes these screens. We show that PMM thereby gains statistical power. PMM can also incorporate weights, for example data can be weighted according to the RNAi quality.
Furthermore, we propose methods for the analysis of genome-wide siRNA screens with small number of replicates. We provide an improved off-target correction method, a threshold for the readout ranking which takes follow-up screens into account and a deconvolution technique of microRNA mimic screens in order to enrich the original screen with more data. Moreover, we show that additional information can be retrieved through the analysis of individual cell features by applying Random Forests.

Finally, we go one step beyond hit detection. We propose an algorithm that exploits off-target effects as multiple knockdowns and reconstructs from them an estimate for the inverse covariance matrix of the pure gene effects. It assumes that the off-target relationships of siRNAs are known. We confirm its efficiency by simulation studies.
Zusammenfassung

Infektionskrankheiten, welche durch Bakterien oder Viren ausgelöst werden, gehören zu den zehn Hauptursachen für den menschlichen Tod. Obwohl im letzten Jahrhundert viele medizinische Behandlungen erfolgreich entwickelt wurden, sind Viren und Bakterien immer noch ein schwerwiegendes Problem für die Volksgesundheit, da die Krankheitserreger resistente Mechanismen gegen existierende Medikamente entwickeln. Im Kampf gegen diese immer resisterter werdenden Eindringlinge sind neue Behandlungsansätze dringend notwendig.

Für die Entwicklung neuer wirksamer Behandlungen ist ein umfangreiches Verständnis über den Eintrittsprozess der Krankheitserreger in die menschliche Zelle essenziell. Groß angelegte RNA Interferenz (RNAi) Experimente über alle Gene hinweg bieten ein mächtiges Werkzeug, um die biologischen Signalwege umfassend für eine Vielfalt zellulärer Prozesse untersuchen zu können. Dazu gehören auch die zellulären Prozesse, welche die Krankheitserreger für das Eindringen in die menschliche Zelle verwenden. Die relativ neue RNAi Technologie erzeugt eine große Datenmenge, die in den meisten Fällen viel Fehlerrauschen enthält und mit ihrer Anzahl Variablen die Anzahl Datenpunkte bei Weitem überschreitet. Das Auswerten solch hochdimensionaler Datensätze ist eine Herausforderung. In dieser Doktorarbeit schlagen wir einige statistische Methoden vor, die aus der großen Anzahl getesteter Genes die einflussreichen Gene herausfiltern, so wie einen Algorithmus, der die partielle Korrelation zwischen den getesteten Genen rekonstruiert.


<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST</td>
<td>basic local alignment search tool</td>
</tr>
<tr>
<td>DAPI</td>
<td>DNA fluorescent stain</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>GFP</td>
<td>green fluorescent protein</td>
</tr>
<tr>
<td>FDR</td>
<td>false discovery rate</td>
</tr>
<tr>
<td>FPR</td>
<td>false positive rate</td>
</tr>
<tr>
<td>Glasso</td>
<td>graphical least absolute shrinkage and selection operator</td>
</tr>
<tr>
<td>GSEA</td>
<td>gene set enrichment analysis</td>
</tr>
<tr>
<td>HCS</td>
<td>high-content screening</td>
</tr>
<tr>
<td>esiRNA</td>
<td>endoribonuclease-prepared siRNAs</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto encyclopedia of genes and genomes</td>
</tr>
<tr>
<td>Lasso</td>
<td>least absolute shrinkage and selection operator</td>
</tr>
<tr>
<td>Lowess</td>
<td>locally weighted scatter plot smoothing</td>
</tr>
<tr>
<td>MIARE</td>
<td>minimum information about an RNAi experiment</td>
</tr>
<tr>
<td>miRNA</td>
<td>micro ribonucleic acid</td>
</tr>
<tr>
<td>MIRZA</td>
<td>biophysical model of miRNA-target interaction</td>
</tr>
<tr>
<td>MOI</td>
<td>multiplicity of infection</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger ribonucleic acid</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>MTT</td>
<td>moderated t-test</td>
</tr>
<tr>
<td>openBIS</td>
<td>open system for managing and distributing biological information</td>
</tr>
<tr>
<td>PMM</td>
<td>parallel mixed model</td>
</tr>
<tr>
<td>RFP</td>
<td>red fluorescent protein</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RNAi</td>
<td>RNA interference</td>
</tr>
<tr>
<td>ROC</td>
<td>receiver operating characteristic</td>
</tr>
<tr>
<td>RSA</td>
<td>redundant siRNA activity analysis</td>
</tr>
<tr>
<td>RT</td>
<td>room temperature</td>
</tr>
<tr>
<td>siRNA</td>
<td>small interfering RNA</td>
</tr>
<tr>
<td>SVM</td>
<td>support vector machine</td>
</tr>
<tr>
<td>TPR</td>
<td>true positive rate</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Infectious Diseases

Infectious diseases caused by viruses, bacteria or parasites are among the top 10 global causes of death. In 2013 about 17% cases of death (9.2 million people) were traced back to infectious diseases (Naghavi et al. 2015). Lower respiratory infections, HIV/AIDS, diarrheal diseases, tuberculosis and malaria represent the leading disease killers worldwide (World Health Organization (WHO) 2004). In low income countries, infectious diseases occur more often due to poor sanitation and insufficient health care systems. However, the number of deaths globally decreased during the last decade since many medical treatments, as antibiotics or anti-viral drugs, have been successfully developed.

Despite the development of medical treatments and the spread of medical applications, bacterial and viral pathogens have acquired resistant mechanisms. Many standard treatments have become ineffective in the fight against these resistant pathogens. Examples of resistant pathogens are HIV (Little et al. 2002), Mycobacterium tuberculosis (Dheda et al. 2014), Salmonella (Threlfall 2002) and Shigella (Ashkenazi et al. 2003). The number of drug resistant and
even multi-drug resistant pathogens has massively increased during the last years and has become a major public concern (World Health Organization (WHO) 2014, Chang et al. 2015). As a consequence, new approaches are required for a successful treatment of infectious diseases.

Drugs are currently designed in such a way that they target a specific mechanism of a pathogen. An alternative approach utilizes a set of proteins in the human host cell as targets in order to prevent the pathogen by its entry into the human cell. This set of proteins needs to include only those that are essential for the cell entry or the replication of the pathogen, but unnecessary for the cell survival. Such a set of proteins increases the therapeutical effect, even more if there is an overlap between different pathogens sharing, for example, a single component in their entry pathway. This alternative approach assumes, as a first step, that the components of the entry process are comprehensively analyzed.

Large-scale RNA interference screens provide a powerful tool for a comprehensive investigation of biological pathways on a wide range of cellular processes (Mohr and Perrimon 2012). Through interfering with the process of mRNA translation, RNA interference enables the association of a certain gene with a specific phenotypical behavior. Some experiments have already been performed to comprehensively identify targets of pathogens, for example for Salmonella (Misselwitz et al. 2011), HIV (Zhou et al. 2008) or influenza virus (Stertz and Shaw 2011).

The systematic RNA interference on thousands of genes generates a large amount of data. Exploiting such large data, in most cases even high-dimensional data ($p >> n$), is challenging since only the application of advanced image analysis and statistical methods enables the extraction of valuable cellular information. Interdisciplinary knowledge is an essential component to succeed in the challenge of generating and analyzing these multi-parametric data.
InfectX is a Switzerland-based consortium, which is part of the framework of SystemsX (the Swiss initiative for systems biology). It has the goal of identifying and analyzing components involved in the pathogen entry into human cells. The pathogen entry is exemplarily studied by a set of four bacterial (*Bartonella henselae*, *Brucella abortus*, *Listeria monocytogenes*, *Salmonella typhimurium* and *Shigella flexneri*) and three viral pathogens (*Adenovirus*, *Rhinovirus*, and *Vaccinia virus*). In order to enable a comparative analysis of similarities and differences between the pathogens, InfectX has a strong emphasis on developing and using unified wet-lab, analysis protocols and work flows. The consortium consists of eleven research groups. The researches in these groups are biologists, each specialized for a different pathogen, computer scientists, responsible for the image analysis of the microscope pictures and the data management, and computational biologists, as well as statisticians, who analyze the data and conceive sound models.

We had the opportunity to be part of the InfectX consortium and studied the phenotypic readouts (e.g. cell count or rate of infection) of the involved pathogens to infer similarities and differences in the process of the pathogen entry into the human cell. Moreover, we aimed to reconstruct the partial correlations between the studied genes for known off-target relationships, which is a first step towards the estimation of a gene network.

### 1.2 Outline

The objective of this thesis is to provide to the biologists powerful statistical methods for studying large-scale data coming from RNA interference experiments, and secondly to contribute to a profound understanding of the pathogen entry into the human cell through the application of the developed statistical methods to the data generated within the InfectX consortium.
Chapter 2 gives a brief introduction to molecular biology. It introduces the biological mechanisms of RNA interference experiments and gives details on the pathogen entry. Moreover, we explain which experiments were performed within the InfectX consortium.

Chapter 3 focuses on a small part of the InfectX data: the RNA interference screens from the kinase libraries. The kinome is an important functional subset of the proteome. The good availability of kinome libraries by different manufacturers and the relatively small number of genes allowed InfectX to perform kinome-wide siRNA screens with a total of eleven independent siRNAs in reasonable time and cost for all the eight different pathogens. This high number of available independent siRNA experiments per gene was a good basis to set up a solid statistical method. We developed a model that simultaneously analyzes the screens of all eight pathogen. We showed that the simultaneous analysis of the readout data increased the statistical power for the detection of influential genes shared between pathogens.

Chapter 4 deals with the genome-wide RNA interference screens performed within the InfectX consortium. In comparison to the kinome-wide data, different challenges emerge. Among these are the high-dimensionality due to the low number of replicates, and the low reproducibility caused by imperfect binding of the siRNAs (off-target effects). We introduce several statistical methods to analyze the readouts of this high-dimensional data, including approaches for the off-target correction. We propose further a threshold for rankings which takes follow-up screens into account, and a deconvolution technique for the data from microRNA mimic screens in order to enrich the original siRNA screen with more data. Moreover, we developed a method for the analysis of data from individual cells in order to evaluate the cell behavior under infection, described by various phenotypical readouts, in more detail.

Chapter 5 goes one step beyond hit detection. We describe an algorithm that is able to reconstruct the partial correlations between the genes from siRNA readout data. Therefore, we assume that the
siRNA off-target relationships are known. The algorithm is based on
the EM-algorithm and the graphical Lasso. Its performance is tested
through a simulation study. The estimated inverse covariance matrix
is a first step towards the estimation of a gene network.

Finally, a brief outlook in chapter 6 summarizes the research pre-
sented in this thesis, mentions further developments and how it could
impact the analysis of RNAi screens or the discovery of new drugs.
Chapter 2

Biological and Experimental Background

This chapter gives a brief introduction to molecular biology. Moreover, it describes all relevant technologies to understand the experiments performed and the data generated by InfectX.

2.1 Molecular Biology

Molecular biology deals with various biological processes in a cell at a molecular level. The ”central dogma of molecular biology” describes the flow from genetic information to proteins (Crick 1958, Crick 1970). The genetic information is mainly situated in the cell nucleus in form of the double-helical DNA molecule, separately stored in chromosomes. A single unit of genetic information in the DNA is called a gene. Through protein synthesis the genes are translated into proteins, which are responsible for nearly every process of cellular life. The protein synthesis consists of two steps: the transcription and the translation (see Figure 2.1). During the transcription phase
a gene is read by the RNA polymerase molecule, which can unwind the double helix, move along the DNA and create an RNA molecule with complementary base sequence. This pre-mRNA leaves after further processing the cell nucleus as mature mRNA molecule. In the cytosol, the translation starts. The mRNA molecule is read on a ribosome in base triplets, known as codons, and mapped to amino acids. In a final step, the amino acids are bound together, modified and folded into a functional protein.

Figure 2.1: An illustration of the protein synthesis in an eukaryotic cell. In the cell nucleus an RNA polymerase attaches to the DNA, unwinds the DNA strands and produces an RNA transcript. This pre-mRNA leaves after further processing the cell nucleus as mature mRNA molecule. The mRNA is transported to a ribosome and there translated into a protein.

The cell can control the expression of a particular gene through microRNAs. MicroRNAs are small (about 22 nucleotides) non-coding RNA molecules. They can bind via base-pairing to a complementary base sequence of an mRNA molecule and cause thereby that a certain mRNA molecule is less (or sometimes even not at all) translated in the ribosome. MicroRNAs are estimated to control the expression of more than 60% of all human protein-coding genes (Friedman et al. 2009). Fire et al. 1998 discovered that these kind of RNA interference can also be provoked by introducing synthetic RNA into cells. This opened new possibilities for the molecular biology to link genes with
2.2 Inference with SiRNAs

Small (or short) interfering RNA (siRNA) is a synthetic double-stranded RNA duplex of 25 nucleotides. SiRNAs are transfected into cells and binds to an mRNA sequence inducing a microRNA-like degradation of the expression of the targeted gene (see Figure 2.2.A). The targeted gene is determined by the antisense strand of the siRNA which is complementary to the base sequence of the target mRNA. Although sequence complementarity appears to be a sound mechanism, scientists discovered that siRNAs bind also to mRNAs with partial complementary base sequence leading to a phenotypical response of the cell which is different to the intended on-target phenotype (see Figure 2.2.B). This sequence-dependent off-targets are primarily caused by the ”seed region” (nucleotide position 2 to 8) (Jackson et al. 2006). Partial complementarity typically occurs several hundred times. There are numerous strategies to reduce the impact of off-target effects on the phenotypic readouts. From the experimental side, it has been shown that a pooling of several inde-
pendent siRNAs designed to target the same gene reduces the number of off-target effects (see Section 3.2.4). Alternatively, the application of a statistical models can help afterwards to extract the signal of the intended knockdown gene from the observed phenotype (see Section 4.2.2).

2.3 Pathogen Entry

Viral pathogens are small particles consisting of nucleic acids surrounded by a protective covering. Bacterial pathogens are slightly more complex particles. Through their well developed cell structure, bacteria are able to perform basic metabolic functions by themselves. All viruses are obligate intracellular pathogens, i.e. they have to enter a host cell for replication. However, only a minority of bacteria are intracellular pathogens. For the invasion of a host cell, bacteria and viruses exploit the cellular functions of the host cell, including receptors, signaling pathways and endocytosis. Due to the complexity of the entry mechanisms, scientists have not yet found any comprehensive model to describe the pathogen uptake with all its molecular interactions. Combining RNAi and the pathogenic infection of cells provides an instrument for the identification of new proteins in the entry pathway.

2.3.1 Viral Entry

In a first step, the viral pathogen binds to the cell surface of the target cell through specific receptors. From there the virus activates a cellular signaling pathway that results in the internalization of the virus (Greber 2002). The uptake is triggered by a variety of different mechanisms, including macropinocytosis, clathrin- and dynamin-dependent endocytosis. For the internalization viruses can use the endocytosis, fuse with the plasma membrane or directly inject their genome into the cytosol. The mechanism is depending on their structure and their replicative niche. Almost all the RNA viruses replicate in the cytosol, whereas the DNA viruses need to be transported to the nucleus for replication. Therefore, most DNA viruses use the
endocytosis since it transports the virus deep into the cytoplasm. Another advantage of the endocytosis is that the virus is protected from destruction through the enclosing endocytic vesicle. The endocytic vesicle signals the DNA virus with a decreasing pH level the location and timing to escape from the vesicle (Marsh and Helenius 2006).

### 2.3.2 Bacterial Entry

Invasive bacterial pathogens induce their entry through phagocytosis by rearrangement of the actin cytoskeletal (Cossart and Sansonetti 2004). Typical entry mechanisms are the zipper mechanism and the trigger mechanism. Bacteria that use the zipper mechanism express surface proteins. These proteins bind to the receptors on the host cell membrane and induce a signaling cascades. The bacterium is internalized as a vacuole by minor actin rearrangement. Contrary to the zipper mechanism, the trigger mechanism uses another signaling cascade which leads to massive actin rearrangement. The membrane ruffles around the bacterium and finally engulfs the bacterium. Another entry mechanisms is the invasion by a mechanism called invasome-mediated uptake (Dehio 2012). Invasome-mediated uptake results in the simultaneous internalization of a large set of bacterial pathogens. The uptake is comprised by massive rearrangements of the actin cytoskeleton. After internalization the bacterial pathogen either replicates in an intracellular vacuole or escapes to the cytosol.

### 2.3.3 Studied Pathogens

Within the InfectX consortium the pathogen entry is exemplarily studied for a set of four bacteria (*Bartonella henselae*, *Brucella abortus*, *Listeria monocytogenes*, *Salmonella typhimurium*, and *Shigella flexneri*) and three viruses (*Adenovirus*, *Rhinovirus*, and *Vaccinia virus*). Table 2.1 lists the known entry mechanisms for these pathogens (see also Figure 2.3). The table shows that the choice of the bacterial and the viral pathogens covers a wide variety of entry mechanisms.
Figure 2.3: **An illustration of the different mechanisms for pathogen entry into human cells.** The left side shows three mechanisms for bacterial entry, the right side two mechanisms for the viral entry. Through invasome-mediated uptake a large set of bacterial pathogens simultaneously enters the cell, with the zipper mechanism a bacterium enters by zippering the plasma membrane around, and with the trigger mechanism a bacterium is engulfed through massive actin rearrangement. Macropinocytosis and clathrin-mediated endocytosis are two trigger mechanism used by viruses to enter the cell.

<table>
<thead>
<tr>
<th>Pathogen</th>
<th>Entry Mechanisms</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Adenovirus</em></td>
<td>clathrin- and dynamin-dependent pathway</td>
<td>Meier et al. 2002</td>
</tr>
<tr>
<td><em>B. abortus</em></td>
<td>zipper-like mechanism</td>
<td>Ackermann et al. 1988</td>
</tr>
<tr>
<td><em>B. henselae</em></td>
<td>invasome-mediated uptake</td>
<td>Dehio et al. 1997, Eicher and Dehio 2012</td>
</tr>
<tr>
<td><em>L. monocytogenes</em></td>
<td>zipper mechanism</td>
<td>Cossart and Sansonetti 2004</td>
</tr>
<tr>
<td><em>S. flexneri</em></td>
<td>trigger mechanism</td>
<td>Patel and Galán 2004</td>
</tr>
<tr>
<td><em>S. typhimurium</em></td>
<td>trigger mechanism</td>
<td>Schlumberger and Hardt 2006</td>
</tr>
<tr>
<td><em>Vaccinia virus</em></td>
<td>macropinocytosis</td>
<td>Mercer and Helenius 2008</td>
</tr>
</tbody>
</table>

Table 2.1: Entry mechanisms of viral and bacterial pathogens studied within the InfectX consortium.


2.4 InfectX Experiments

InfectX generated various large-scale siRNA experiments for a comprehensive analysis of the entry mechanisms of bacterial and viral pathogens into human cells. In a first step the RNA interference was performed on a kinome-wide scale, in a second step on a genome-wide scale and in a last step several genes were selected and validated through follow-up experiments. These validation experiments were enriched by additional screens with esiRNAs and microRNAs mimics. Each RNAi screen was carried out in several labs, each specialized for a certain bacterial or viral pathogen. In order to obtain reproducible results, InfectX had a strong emphasis on the development of common protocols for cell culture, RNAi transfection, pathogen infection, microscopy and image analysis. The InfectX work flow is illustrated in Figure 2.4.

2.4.1 RNAi Screens

The siRNA libraries were all arrayed on 384-well plates using HeLa (ATCC CCL-2), the oldest and widely used human cell line. The culturing of the cells was performed at 37°C, under 5% CO₂ atmosphere. The libraries for the RNA interference were obtained from three different manufactures (Ambion, Dharmaco and Qiagen) in order to avoid a potential bias introduced by the siRNA design of a certain manufacturer. Further, InfectX used several siRNA sequences, as well as pooled and unpooled siRNAs, to account for the off-target effects. Screens were repeated 1-8 times depending on the library and the pathogen. Exact numbers are listed in Table 3.1 for the kinome-wide screen and in Table 4.1 for the genome-wide screens.

At 72h posttransfection the HeLa cells were infected with a pathogen. This infection phase was designed by a pathogen-specific protocol based on two criteria. These two criteria should ensure comparable phenotypes in terms of infection rate, as well as a reliable detection of genes that increase or decrease the infection rate through
Figure 2.4: **InfectX workflow.** InfectX transfected HeLa cells with several pooled as unpooled siRNA libraries from the three manufactures Ambion, Dharmacon and Qiagen. At 72h posttransfection the HeLA cells were infected with a pathogen. All screened plates were fixed, stained and then imaged by automated microscopy. InfectX performed all imaging with the Molecular Devices ImageXPress microscopes with 10x magnification on 3-4 channels. Image analysis consisted of image shading correction, object segmentation (nuclei, perinuclei, cells and Voronoi cells), feature extraction (typically 300 features per cell) and infection scoring (with a decision tree algorithm). The classification into infected and not infected cells was then aggregated to an infection rate or cell count per well.

their knockdown. On the one hand we wanted to reach for all eight pathogens an infection rate close to 30-50% in a well, and on the other hand we aimed to reach a cell number of around 1500 cells per well in order to leave some empty space between the cell colonies. Due to the distinct infection biology of the different pathogens these criteria required adaptations. For *B. abortus* the infection rate typically remained low (close to 5%) and for *B. henselae* the infection rate was typically high (close to 90%). Table 3.2 summarizes the key parameters that varied between pathogens, including cell number, multiplicity of infection (MOI) and infection times. In a last step, the HeLa cells were fixed and stained for DNA (using DAPI or Hoechst), F-actin (using fluorescent phalloidin) and infection specific markers (using immunofluorescence). The detailed protocols can be found in section 3.A.3.
The number of possible known error types is very large for complex high-throughput and high-content data. Therefore, InfectX created procedures to tackle possible errors, for example the failure of siRNA transfection or cytotoxicity of siRNAs. For the RNAi screens, InfectX performed three kind of controls: positive, negative and mock. Positive controls were used to confirm expectations. For example, cells were transfected with siRNAs for which their targeted gene is known to be essential for the cell cycle or the cell survival. Consequently, we expected the cells to die after this transfection. As negative controls InfectX used, for example, siRNAs with a random seed sequence, having thus no specific gene as target and resulting in an unaffected phenotypic readout. At last mock controls refer to wells where only the pathogen was added but no transfection with an siRNA was conducted beforehand.

2.4.2 From Images to Data

All screened plates were imaged by automated microscopy. The imaging was performed with ImageXpress micro (IXM) HCS microscopes from Molecular Devices at the University of Basel, as well as at the Light Microscopy and Screening Center of ETH Zürich. Wells were divided into 9 sites (3 × 3 grid) with no spacing and no overlap (see Section 2.A for more details). The image analysis started with scaling and shading correction in order to address illumination and sensor inhomogeneities. Then the following objects were segmented using CellProfiler (Carpenter et al. 2006):

1. **Nucleus:** The cell nucleus was extracted using the DNA stain values. The morphology of the nucleus and the uniform distribution of DNA inside the nucleus allowed a reliable detection.

2. **PeriNucleus:** The perinuclear object refers to the intracellular area next to the nucleus. It was segmented through the extension of the nuclear border by 8 pixels and the deletion of the nuclear area.
3. **CellBody**: The segmentation of a cell was based on the actin intensity in the cytoplasm. Irregular shapes and inhomogeneities in the actin distribution made this a challenging task.

4. **VoronoiCell**: A non-actin based cell body object is the Voronoi cell. It was constructed by adding a maximum of 24 pixels around the nuclear border. The extension stops as soon as the border intersects with the border of a neighboring Voronoi cell.

5. **Pathogens**: Due to the heterogeneity of the infection pattern, sometimes single pathogens and sometimes clusters of pathogens were detected. The detection was mainly based on the GFP signal.

For the segmented objects the following features were analyzed: intensity, location, texture and neighborhood relationships (see Figure 2.5). Intensity, area and texture measurements were calculated by summary statistics over the different imaging channels. The location was determined in pixel distances. The neighborhood relationship was measured within a two pixel distance between cell body objects (see Section 2.C.1 for more details).

**Figure 2.5:** An illustration of the features extracted from an individual cell. (A) The following objects were segmented for each individual cell with the help of the actin stain, DNA stain and pathogen immunofluorescence: nucleus, peri nucleus (extension of the nuclear border by 8 pixel), Voronoi cell (extension of the nuclear border by a maximum of 24 pixels), complete cell body, and pathogens. (B) The segmented objects were analyzed for intensity, location and texture by using various summary statistics. The total number of features per cell is around 300.
2.4.3 Infection Scoring

The features of an individual cell served as bases for the automatic detection of infected cells in the microscope images. The detection was performed by pathogen-specific models using a decision tree or a support vector machine (see Section 2.C.2 for more details). For each pathogen only a small number of features was selected (typically 2-5 features). For example, for *B. abortus* important features were the GFP mean intensities across the objects Nuclei, PeriNuclei and Cell body, which capture the typical infection pattern of large micro-colonies spread throughout the cell, or for *S. typhimurium* an important feature was the GFP signal over the whole cell body (see Section 2.C.3 for more details). The choice of the thresholds was adjusted on a plate-by-plate basis in order to take the staining quality, cell vitality or microscope illumination into account.

All classification methods yielded a binary response on cell level: cell is infected or cell is not infected. The binary infection phenotype allowed the definition of an infection index per well, which enables cross-pathogen comparison after standard normalization:

\[
\text{Infection Index} = \frac{\#\text{Infected Cells}}{\#\text{Cells}}.
\]

This infection score, as well as the number of cells per well are used for almost all of our statistical analysis in the following Chapters 3, 4 and 5.
Appendix Chapter 2

2.A Microscopy

All screened plates were imaged by automated microscopy using a robotic plate handling to load and unload plates (Thermo Scientific). The imaging was performed with ImageXpress micro (IXM) HCS microscopes from Molecular Devices at the University of Basel, as well as at the Light Microscopy and Screening Center of ETH Zürich. The settings for both microscopes were equal. Wells were divided into 9 sites (3 × 3 grid) with no spacing and no overlap, and then imaged with 12 bit dynamic range and laser-based focusing. The Z-Offset for Focus was manually selected and AutoExpose was used to get a good exposure time. The exposure time was manually corrected, whenever it was necessary, to ensure a wide dynamic range with low overexposure.

2.B Computational Infrastructure

Computational infrastructure is one of the key aspects in order to achieve reproducible high-content screening (HCS) results. HCS requires an infrastructure with multiple terabytes of storage (a single plate yields around 10000 images ∼ 20 GB) for the experimental data, with clusters for efficiently running computationally intensive tasks, with storage for the analysis results and a possibility to share the data and results. An open-source work flow management solution called iBRAIN2 (Rouilly et al. 2012, http://ibrain2.sourceforge.net/) was developed for the InfectX consortium in order to address all requirements of data storage and analysis. IBRAIN2 was connected with openBIS, a data management solution (Bauch et al. 2011). OpenBIS offered an extended support for screening metadata (e.g. siRNA library) and screening results (e.g. microscope images or features from single cell analysis). This computational setup enabled the parallel analysis of HCS data sets on high-performance computing clusters.
Image analysis, as well as pre-processing of the data was performed on a Linux-based cluster of heterogeneous multicore nodes (x86_64 architecture with 2GB RAM per process) running Sun Grid Engine on CentOS Linux. Data was stored on an NFS-mounted IBM SONAS storage system.

The InfectX consortium had a strong emphasis on the reproducibility of their results. Therefore, all computational results of the screening data were linked with metadata referring to the used methods and settings. Moreover, new methods were extensively tested by several people.

2.C Image Analysis

2.C.1 Object Detection and Feature Extraction

From all microscope images pixel intensities were extracted and scaled to the [0, 1]-interval. In a next step the microscope images were corrected for shading. Therefore, images were overlaid and the foreground and background intensities were estimated per image pixel. The shading model was adjusted for the following factors: microscope, lamp, channel, pathogen and assay.

CellProfiler, an open-source modular platform for cellular image analysis (Carpenter et al. 2006), was used for the segmentation of cellular objects and the feature extraction. The nucleus object was detected by OTSU’s method (CellProfiler module IdentifyPrimAutomatic). The peri-nuclear object was segmented with help of the CellProfiler modules ExpandOrShrink and IdentifyTertiary. The cell body was constructed from the intensity gradient changes of the actin channel (Propagation method, CellProfiler module IdentifySecondaryInformed). At last the voronoi cell was segmented by extending the nucleus object by a maximum of 25 pixels (CellProfiler module ExpandOrShrink). Pathogens objects were additionally detected for suitable assays based on OTSU method or on wavelets.
Subsequently to the segmentation, measurements were taken on the individual image areas. They were analyzed for intensity, location, texture and neighborhood relationships. Intensity, area and texture measurements were calculated through summary statistics over the different imaging channels. The location was determined in pixel distances. The neighborhood relationship was measured within a two pixel distance between cell body objects. The extracted features from individual cells were named by the following rule:

<object>.<group>_<measurement>_<channel>

An example is VoronoiCells.Intensity_Max_CorrPathogen which corresponds to the maximal intensity of the GFP signal within the segmented object VoronoiCell. All features were stored in the open-BIS database and linked to their corresponding microscope image.

2.C.2 Methods for Infection Scoring

The InfectX consortium used several different algorithms to detect the binary infection phenotypes: decision tree, support vector machine and a segmentation based algorithm. The algorithms are described below. All algorithms yielded a binary infection score (cell is infected or not infected) as response. For quality control the InfectX consortium applied per pathogen at least two of the infection scoring algorithms. The approach for the detection of infected cells was established pathogen-specific to account for the distinct infection pattern (see Section 2.C.3).

Decision Tree Infection Scoring

Using the knowledge of the biologists $N$ features were selected which were most sensitive to the infection phenotype. These selected features were evaluated on a training data in a decision tree in order to find suitable thresholds. The training data were produced by expert labeling. Figure 2.6 shows examples of for infection scoring with the decision tree algorithm. Thresholds were chosen plate-specific
**Support Vector Machine Infection Scoring**

Classification through a Support Vector Machine (SVM) finds a hyperplane that best separates the data into two groups for a given set of training data. SVM yielded good results for pathogens that had a clear binary infection phenotype (e.g. *S. typhimurium* or *Vaccinia virus*). However, the results were not so good for pathogens that had relatively continuous infections levels (e.g. *L. monocytogenes*). The SVM model was fitted with CellClassifier (Rämö et al. 2009).

**Segmentation based Infection Scoring**

Pathogen object segmentation directly leads to a binary infection scoring for each cell. A cell was defined as *infected* if a pathogen object was detected within the cell. If the pathogen object was overlapping with two cells, then it was assigned to the cell with which it had more overlap.
Infection Scoring for Each Pathogen

**Adenovirus** infection shows a dispersed GFP signal across the whole cell body. Since the strength of the signal strictly depends on the amount of virus added to the cells, the GFP intensity was normalized by measuring the GFP intensity in all segmented cell objects. There was no segmentation of the viral objects.

**B. henselae** infection appears in invasomes. Invasomes are characterized by an actin surrounded membrane. Therefore, invasome object detection was performed by matching idealized invasomes of varying size to the actin channel. A cell was classified as infected if the cell contains one or more invasomes.

**B. abortus** infection appears as wide spread micro colonies across the whole cell body. The GFP intensity was measured in all segmented cell objects. There was no segmentation of the bacterial objects.

**L. monocytogenes** infection shows a dispersed Cy3 signal across the whole cell body. Since the strength of the signal depends on the amount of bacteria in the cell, the Cy3 intensity was measured in all segmented cell objects. There was no segmentation of the bacterial objects.

**Rhinovirus** infection appears as small cytoplasmic clusters of the viral replication sites. They are detected by the monoclonal antibody 16-7, and a secondary anti-mouse IgG conjugated to Alexa488. Wavelet-based object detection was used to segment the pathogen. The fluorescence intensity was measured in the viral object.

**S. flexneri** infection appears in form of one or more accumulations of micro colonies. They are mostly localized in the peri nuclear area. Bacteria objects were segmented using Otsu’s method based on the fluorescent marker. The RFP intensity was measured in the bacterial objects.
S. typhimurium infection appears as small GFP dots across the whole cell body. Wavelet-based object detection was used to segment the bacterial objects and the GFP intensity was measured in the segmented bacterial objects.

Vaccinia virus infection appears as dispersed GFP signal across the whole cell body. The GFP intensity was measured in all segmented cell objects. There was no segmentation of the viral objects.
Chapter 3

Simultaneous Analysis of large-scale RNAi Screens\(^1\)

We propose the Parallel Mixed Model (PMM) approach that simultaneously analyzes several non-identical screens performed with the same RNAi libraries. We show that PMM gains statistical power for hit detection due to parallel screening. PMM allows incorporating siRNA weights. These weights can be assigned according to available a-priori information on the RNAi quality. Moreover, PMM estimates a sharedness score that indicates whether a gene is a generic or specific regulator among the studied pathogens. By fitting the PMM model to the InfectX kinome-wide RNAi screening data, we found several new hits for most of the pathogens studied. Our results show that parallel RNAi screening can improve the results of individual screens.

\(^1\)This chapter is a slightly modified version of Rämo, P and Drewek, A et al. 2014. Pauli Rämo and Anna Drewek are shared first authors and wrote this paper together. The contributions of Anna Drewek are the development of PMM (including the sharedness score), the simulation and resampling studies for testing the goodness of PMM, as well as the enrichment analysis.


3.1 Background

Large-scale RNAi screening is a widely used technology to knock-down expressions of genes and study their protein function in a biological process of interest (Conrad and Gerlich 2010, Mohr et al. 2010, Mohr and Perrimon 2012, Simpson et al. 2012, Elbashir et al. 2001). In several published studies in the field of infection biology, cells perturbed with siRNAs were exposed to pathogens. Differences in phenotypic outcomes were measured under knockdown in order to identify the genes involved in successful infection or to develop functional models of host signaling and trafficking pathway (Mercer et al. 2012, Stertz and Shaw 2011, Misselwitz et al. 2011, Snijder et al. 2012).

RNAi libraries are mostly sold in formats containing enough material for numerous large-scale screens. Therefore, several large-scale siRNA screens are typically performed using the same libraries within a unit such as a university or company in order to optimize material costs and to simplify plate handling. However, parallel screens are typically performed and analyzed separately without common protocols or analysis pipelines. Therefore, comparing results between the screens is challenging. Co-operative efforts, such as assays using common key parameters for imaging and data analyses, could enable more comparable results. By this the parallel information could improve the results of each individual screen. In the field of RNAi screening, there has been progress in relation to the standardization of data publication formats, in particular in the context of the Minimum Information About an RNAi Experiment (MIARE, http://miare.sourceforge.net) and GenomeRNAi (Schmidt et al. 2013). Nevertheless, the provided meta data information and data analysis approaches are often diverse so that data comparison between the screens from different laboratories is very difficult.

A common concern of large-scale RNAi screens are the poor reproducibility rates. They are mainly caused by strong off-target effects from the siRNAs (Pache et al. 2011, Sigoillot and King 2011,
3.2 Results

3.2.1 Kinome-Wide RNAi Screening for Infection

The InfectX consortium generated kinome-wide siRNA screens for five bacterial and three viral pathogens (see Figure 3.1). The screens were conducted in a highly parallel manner under one common protocol for all eight pathogens. The set of 826 targeted genes comprised almost the whole kinome, plus selected kinome-associated genes. Each gene was targeted by a total of eleven independent siRNAs coming from three manufactures: Ambion (Silencer Select) with 3 siRNAs per gene, Qiagen (Human Kinase siRNA Set V4.1) with 4 siRNAs
Chapter 3: Simultaneous Analysis of large-scale RNAi Screens

Figure 3.1: Overview of InfectX high-content datasets, image analysis, and the Parallel Mixed Model (PMM). (A) The figure shows example images of the different pathogens after siRNA transfection and the infection phase. The arrows indicate typical infectious phenotypes for each pathogen. The list shows an example of three single cell features that we extracted to identify infected cells for *L. monocytogenes*. The scale bar has a length of 50 µm. (B) For each selected feature, we defined the optimal threshold that separated best between infected and uninfected cells. We used the thresholds in a Decision Tree algorithm to classify between infected (green) and non-infected cells (red). We optimized this procedure for each pathogen separately. (C) For each well in a 384-well assay plate, we calculated the infection index by dividing the number of infected cells (green) by the total number of cells (green and red). (D) The figure shows a schematic representation of the input data for the statistical analysis. Each point represents the average infection index over all its replicate wells (i.e. wells with the same siRNA set targeting the same gene and pathogen). (E) The PMM fits first an overall effect $a_g$ to all data of one gene $g$ and secondly a correction $b_{pg}$ of the overall effect $a_g$ within every pathogen in order to obtain to an pathogen and gene specific effect $c_{pg}$. The different sizes of the data points refer to weights $w_s$ which can be incorporated in the PMM to depict the quality of the siRNA. (F) The figure represents the final output of the PMM. The model estimates gene effects $c_{pg}$ for each gene and pathogen and provides a local False Discovery Rates $q_{pg}$ for each estimate.
per gene and Dharmacon (Human ON-TARGETpIus) with 4 siRNAs per gene. Additionally, the InfectX consortium performed screens where each kinase was targeted with a pool of the 4 Dharmacon siRNAs (Human ON-TARGETpIus SMARTpool). Depending on the pathogen and library, we independently repeated the screens one to six times (see Table 3.1 in Appendix 3.A.1). To obtain an optimal dynamic range of infectivity, the pathogen dose and entry time was chosen pathogen specific (see Table 3.2 in Appendix 4.A).

As described in Section 2.4.2 each well was imaged by automated microscopy. Through pathogen-specific image analysis the cells were then classified as infected or uninfected in each well in order to obtain a infection rate per well (infection index) (see Figure 3.1.B–C). We alleviated possible batch effects, dependencies to the population context, and further experimental confounders by data normalization (see Section 3.4.1 for more details) (Malo et al. 2006, Carpenter et al. 2006, Birmingham et al. 2009, Siebourg et al. 2012, Snijder and Pelkmans 2011, Rouilly et al. 2012). We performed all analyses presented in this chapter with the normalized infection index readout, unless otherwise stated.

### 3.2.2 Data Reproducibility

Our data confirmed the reported (Marine et al. 2012) low reproducibility rates of siRNA data originating from off-target effects. The normalized infection indices of two different siRNA sets targeting the same genes showed a Pearson correlation coefficient $R$ between 0 and 0.2 depending on the screens (see Figure 3.2.B). Adding independent siRNAs to the screen yielded an increase in the Pearson correlation coefficients, but the correlation still stayed at an unsatisfactory level, even with an average over six independent siRNAs targeting each gene ($R$ was between 0.1 and 0.4). In contrast, replicate screens (screens performed using the same protocols and siRNA set, but performed at a different time) were reproducible ($R$ was between 0.5 and 0.9) (see Figure 3.2.A). For practical reasons performing screens in duplicates seems sufficient assuming a similar assay quality as ours.
Figure 3.2: Using more siRNAs adds power and yields reproducible results. (A) The three boxplots show the Pearson correlation coefficients $R$ between the screens which were performed using the same siRNA set. The numbers 1 to 3 correspond to the total number of replicate screens that were averaged and compared to another distinct set of replicate screens, averaged over the same number. To quantify the variation, we resampled the replicate screens up to 500 times. (B) The set of six boxplots shows the Pearson correlation coefficients of the averaged readouts from 1 to 6 siRNA sets. (C) All three scatter plots show examples of the infection index from the Adenovirus screen. The first scatter plot illustrates the correlation between two identical siRNA replicates. The second shows the correlation between two different siRNAs. The third plot shows the correlation between two identical siRNAs. (D) The set of six boxplots shows the Pearson correlation coefficients $R$ between the screens which were performed using the same siRNA set. The numbers 1 to 3 correspond to the total number of replicate screens that were averaged and compared to another distinct set of replicate screens, averaged over the same number. (E) The three boxplots show the Pearson correlation coefficients $R$ between the screens which were performed using the same siRNA set. The numbers 1 to 3 correspond to the total number of replicate screens that were averaged and compared to another distinct set of replicate screens, averaged over the same number.
Having more replicates does not improve the data to a great extent (see Figure 3.2.A). On the other hand, performing screens at least in duplicates is necessary for quality control and performing only single screens is therefore not recommendable. The cell number readouts (see Figure 3.2.D-F) showed qualitatively similar results for data reproducibility. In summary, the main error noise was caused in our siRNA screens by varying specificity of siRNAs and not by technical variability of the screens.

3.2.3 Parallel Mixed Model (PMM)

Assuming that the sources of variability between different siRNAs targeting the same gene are statistically independent, we can benefit from the fact that the true signal is enhanced by using more siRNAs targeting each individual genes (Bassik et al. 2013) (see Figure 3.2.B). In order to increase the statistical power of individual siRNA screens, we performed screens with 11 siRNAs (and one pool of siRNAs) targeting each gene. Moreover, when using the parallel structure in the data and combining data points from all pathogen screens together, we reached 8x12=96 data points for every gene (averaging over the replicate screens). We propose the Parallel Mixed Model (PMM) as a suitable approach to model the distribution of the siRNA readouts using all data together, including all available siRNAs and pathogen screens.

PMM is composed of a linear mixed model and an assessment of the local False Discovery Rate (FDR) (see Figure 3.1.E–F). The linear mixed model is an extension of the ordinary linear model by random effects (Pinheiro and Bates 2000). In particular, random effects are not determined by fixed coefficients, but rather by Gaussian distributions. Therefore, we can incorporate the variation among the siRNAs in form of random effects and estimate all effects for different pathogens simultaneously. To be more precise, the linear mixed model consists of a fixed effect $\mu_p$ for pathogen $p$ and two random effects $a_g$ for gene $g$ and $b_{pg}$ as a correction term for gene $g$ within
Figure 3.3: Parallel screens add power to find more shared hits. (A) We resampled 1000 times data with varying number of siRNAs and pathogens. We calculated the rank of MET for *L. monocytogenes* by PMM (resp. by MTT for the case of one pathogen). The color corresponds to the variation of the observed ranks. The star indicates the boxplots that are significantly different from zero (one sample t-test $p<0.05$). (B) The figure shows the same resampling study as in (A), but now with MTOR for *Vaccinia virus*. The boxplot shows that MTOR is a shared significant hit for several pathogens. (C) The figure shows the same resampling study as in (A), but now with non-hit ALK for *B. abortus* for control.
3.2 Results

pathogen $p$:

$$y_{pgs} = \mu_p + a_g + b_{pg} + \varepsilon_{pgs},$$

where $y_{pgs}$ denotes the readout (for example the normalized infection index of a well) of pathogen $p$ and gene $g$ knocked-down with siRNA s and $\varepsilon_{pgs}$ denotes the unobserved error term. The sum of two random effects $a_g$ and $b_{pg}$ describes the total effect of the siRNAs within pathogen $p$. We define the estimated effect $c_{pg}$ for gene $g$ within pathogen $p$ as

$$c_{pg} = a_g + b_{pg}.$$

A positive estimated $c_{pg}$ effect means that the infection level was enhanced if the corresponding gene $g$ is knocked down. A negative effect means that the infection level was reduced. To distinguish hit genes, PMM provides as second step an estimate $q_{pg}$ of the local False Discovery Rate (FDR). We computed the local False Discovery Rate using the approach presented in Efron 2010. We assigned the local False Discovery Rate to every gene and interpreted it as the probability describing how likely the corresponding gene is a false discovery (see Section 3.4.3 for more details). The implementation of the PMM is published as PMM R-package on Bioconductor (Drewek 2015).

As a first verification for the increase in power by simultaneously using the parallel screening structure, we resampled datasets, each consisting of a fixed number of siRNAs and pathogens, and fitted the PMM, respectively the Moderated T-Test (MTT) (Smyth 2004) for the case of one pathogen (see Section 3.4.6 for more details). We evaluated the mean and variation (i.e. stability) of the ranks in the ordered lists of genes based on their estimated $c_{pg}$ values over 1000 resampling runs for MET (a known effector gene for L. monocytogenes (Pizarro-Cerdá et al. 2012)), MTOR (a role of MTOR in the infection pathways of several pathogens has already been established (Mercer et al. 2012, Sivan et al. 2013, Tattoli et al. 2012)) and a non-hit ALK as control (see Figure 3.3). The results showed, in particular in the case of MTOR, that the rank and its stability improved by simultaneously using more siRNAs and pathogens. In the case of MET the use of parallel screens did not cause an increase in statistical power, since MET was a hit for L. monocytogenes only. However, for MET
there was no reduction of statistical power either. These examples already indicated that the parallel screening structure and PMM can be used to more reliably discover expected effector genes even in the case where only a fraction of effector genes is shared between the screens. The increase in power was especially high in the case of low number of replicates.

3.2.4 Analysis of SiRNA Libraries

PMM allows the assignment of weights to each siRNA (see Section 3.4.3 for more details). With weighting, we can assign more power to siRNAs that are estimated to have little off-target effects and strong knock-down efficiency. Within this study, we weighted siRNAs according to the reproducibility in terms of correlation of their corresponding library to other libraries (see Figure 3.4.A). There are several potential other ways how weights could be determined.

We cross-validated different libraries to each other by fixing one library manufacturer (training set) at a time (see Figure 3.4.A). We averaged phenotypic readouts from siRNAs targeting the same gene in the training set. Through this we obtained reference gene readouts basing on the fact that the true signal is enhanced by the average of independent siRNAs. We then compared single siRNA readouts of the remaining two library manufactures (test set) to the reference gene readouts. The Pearson correlation coefficients of the test sets enable to quantify which of the two test manufacturers produces more reproducible results. By repeating the procedure for all manufacturers as the training set we could order the manufacturers in terms of their reproducibility performance. We used both infection index and cell number readouts for this analysis.

Our results showed that the pooled Dharmacon library performed the best with respect to our phenotypic readouts. The pooled library was followed by the unpooled libraries of Ambion, Dharmacon, and Qiagen in this order. However, there were no statistically significant
Figure 3.4: **Statistics on used siRNA libraries and hits.** (A) We weighted siRNAs based on their library quality. Each vertical compartment in the plot corresponds to a training set of siRNAs. Each boxplot corresponds to a test set of single siRNAs from different manufacturers (except “Dharm. siRNA mean” which is the average of 4 Dharmacon unpooled siRNAs). Y-axis refers to Pearson correlation coefficients R between the training and test sets. A star corresponds to significant differences in the correlation coefficients (Mann-Whitney-U-test p < 0.05) between pairs of manufacturers. We used all screens, infection index, and cell number well readouts in this analysis. We used the results to assign siRNA weights to siRNAs from different library manufacturers as shown below the plot. (B) The histogram shows the FDR values from PMM using the infection index readouts as input. The red line shows the FDR-threshold of 0.4. (C) The bar shows number of up and down hits for different pathogens. (D) The bar plot shows the number of hit genes against the number of pathogens that share the hits.
differences (Wilcoxon rank-sum test \( p < 0.05 \)) between Dhharmacon pooled and Ambion single, and Dhharmacon single and Qiagen single siRNA data reproducibility. In addition, the data showed that the averaged single siRNAs of Dhharmacon performed at most as good as the single pooled siRNA consisting of the same siRNAs. This indicated that for most screening purposes, it is more practical to use the pooled library instead of several unpooled libraries. This result of better performance of pooled libraries compared to averaged single siRNA libraries is in contradiction with what has been reported in Collinet et al. 2010. However, good quality single siRNA libraries (such as Ambion Silencer Select) performed nearly as well as pooled libraries of less good single siRNAs (in our case Dhharmacon SMARTpool). Following the results of the library analysis, we assigned a higher weight to Dhharmacon Pooled and Ambion libraries (weight 2) than to the unpooled libraries Dhharmacon and Qiagen (weight 1). PMM benefitted from the assigned library weights. The residual standard error of the linear mixed model reduced from 0.87 to 0.83.

3.2.5 Sharedness of Detected Significant Genes

By fitting PMM to our data, we found a left tailed local False Discovery Rate distribution, ending with a set of 48 different genes that reached the threshold of 0.4 (see Figure 3.4.B and Figure 3.5.A). We selected threshold 0.4 as a reasonable hit threshold for this study since the difference was small compared to the set of hits with the commonly used threshold 0.2 and 40% false-positive rate was still acceptable for follow-up experiments. The number of up and down hits varied between the pathogens (see Figure 3.4.C). 80% of hits were unique and 20% of hits were shared between two or more studied pathogens using the FDR threshold 0.4 (see Figure 3.4.D). This provided a rough estimate that about 20% of genes gained statistical power from the parallel analysis using the PMM with our data.

To quantify the hits according to their level of being shared between screens, we developed the following “sharedness score” \( s_g \):
Figure 3.5: **Summary of screening hits for all pathogens.** (A) The heat map shows all genes which were significant (FDR < 0.4) for at least one pathogen. We ordered the genes by their averaged $c_{pg}$-values over all pathogens. The colors correspond to the estimated $c_{pg}$-values. The black squares indicate significant hits and the green squares high-light the strongest down and up hits for each pathogen. The rightmost column shows the sharedness scores for each gene. (B) The network shows the hit genes (FDR < 0.4 for at least one pathogen) and their direct neighbors that had connections between kinases in the STRING database (version 9.0). The edges are functional interactions with edge threshold 850. We removed genes that were not connected to any other gene from the network. Each node consists of a colored pie chart, in which each piece corresponds to a pathogen.
\[
s_g = \frac{1}{2} \left( \left( 1 - \frac{1}{n} \sum_{p=1}^{P} (q_{pg}) \right) + \left( \frac{1}{P} \sum_{p=1}^{P} 1 \{ q_{pg} < 1 \} \right) \right).
\]

Here \( P \) is the total number of pathogens (8 in our case). The sharedness score is a combination of two quantities and is independent from the FDR-threshold. The first part defines the shift away from 1 and the second part describes how many pathogens support this shift (proportion of \( q_{pg} < 1 \)). The score returns a value between 0 and 1 for each gene. Score 0 indicates that a gene is not shared among the pathogens and score 1 indicates that the gene is significant among all pathogens (see Figure 3.5.A). Since the sharedness score takes only the strength of a gene and not the directionality into account, a gene can be also highly shared if it inhibits in one pathogen and enhances the infection in another pathogen. Therefore, a gene shared between pathogens should be interpreted as being involved in the entry process of these pathogens.

### 3.2.6 Comparison to Existing Ranking Methods

In order to validate the PMM approach and its results we compared it to other existing hit ranking methods and performed different kind of statistical tests. As reference methods we selected the Moderated T-Test (MTT) (Smyth 2004) and Redundant SiRNA Activity (RSA) (König et al. 2007) which are both commonly used in high-throughput RNAi screening (see Section 3.4.4 and Section 3.4.5 for more details). We could not apply other widely used hit ranking methods, such as Strictly Standardized Mean Difference (SSMD) (Zhang et al. 2007) or percent inhibition (Birmingham et al. 2009) since many of our pathogen screens did not have effective and reliable positive and negative control wells.

As a first test, we analyzed the stability of the obtained gene rankings with respect to the estimated \( c_{pg} \) values (Siebourg et al. 2012, Meinshausen and Bühlmann 2010). We resampled with replacement 1000 datasets (12 siRNAs randomly selected with replacement for each gene) and calculated the number of genes that appear with high
3.2 Results

Figure 3.6: **Performance statistics of hit ranking methods.** (A) The figure shows stability curves using the three different methods (PMM, MTT and RSA). The y-axis denotes the number of genes that were found with probability higher than 0.7 (dashed lines) and 0.9 (solid lines) in the top k (x-axis) of the list of ranked genes. The curves show the average over all eight pathogens. (B) The figure shows hit overlaps of cross-validated siRNA sets between the set of 10 unpooled siRNA libraries and the remaining siRNA library using the three tested gene ranking methods as a function of hit threshold k. The curves show the average over all eight pathogens. (C) The figure shows ROC-curves for PMM, MTT and RSA applied on simulated data containing only hits that were shared between all pathogens. The dashed and solid lines indicate whether the hits were generated by a low or high shift away from zero. The PMM method outperformed the reference hit detection method. (D) The figure shows ROC-curves for PMM, MTT and RSA applied on simulated data containing only unique hits for all pathogens. PMM and Moderated T-Test performed equally well. (E) The figure shows ROC-curves for simulated data with a mixed hit structure (unique and shared hits). The PMM method outperformed the other reference hit detection method.
probability (prob $> 0.9$ and prob $> 0.7$) in the top of the ordered lists of genes based on their estimated $c_{pg}$ values (see Section 3.4.7 for more details). This measure of stability showed similar results for PMM and the reference methods MTT and RSA (see Figure 3.6.A).

To mimic primary and validation screening setup and to study hit reproducibilities of the gene ranking methods we performed a leave-one-out cross-validation experiment. We used the siRNAs of unpooled libraries (11 in total) and left one siRNA set at a time away. We ran PMM, MTT, and RSA on the data sets consisting of 10 individual siRNAs and compared the resulting gene ranking to the ranked gene list of the remaining siRNA set (see Section 3.4.8 for details). Figure 3.6.B. shows the overlap of hits in the test and training set as a function of hit threshold $k$. PMM performed the best indicating that the hits found by PMM are more reproducible by an independent siRNA screen than the hits found by the other methods.

In order to further estimate the hit-calling performance for different methods we performed data simulation with a-priori known hit structure. Data simulation was required since reliable ground truth hits are generally not available for the studied biological systems. We simulated data by generating 1000 Gaussian distributed screens for each pathogen with four siRNAs. We selected four siRNAs since it makes up a realistic screening approach referring to cost and time constraints. We incorporated hits in each simulated screen by randomly selecting 10% of the genes and shifting them away from zero. We distinguished between three types of simulated data. In the first case the hits were different for each pathogen (unique hits only) and in the second case all hits were shared between the pathogens. The third case is probably the most realistic scenario containing both unique and shared hits to a varying degree (see Section 3.4.9 for more details). We then applied PMM, MTT, and RSA to the simulated data and evaluated them by Receiver Operating Characteristic (ROC) curves (with false positive versus true positive rates plotted for diverse FDR- threshold; see Figure 3.6.C–E). The results showed that PMM performed the best especially in the case of shared hits. For the case of unique hits PMM and MTT exhibited about the same
3.2 Results

Figure 3.7: Influence of Parallelism and the differences to other hit scoring methods. (A) Y-axis shows the PMM gene ranking for *L. monocytogenes*. X-axis is the same, but we randomized the other 7 parallel assays. The colors correspond to hit genes (FDR < 0.4). Parallelism yielded only a slight effect on the ranking, but added genes to the list of significant hit genes. (B) The scatter plot shows PMM hit ranking (y-axis) compared to the MTT hit ranking (x-axis) for *L. monocytogenes*. The dot size corresponds to the sharedness score of each gene. Some genes with high sharedness scores gained statistical power.
performance while RSA performed the worst. As expected, the ROC curves got better for all methods, when the hit genes were incorporated with a larger shift away from zero.

We also studied how simultaneous modeling affects the ranking of genes in individual screens using PMM. We performed a test where we selected a pathogen and created two datasets. The first dataset was the full data without any changes and the second dataset had the original data for the selected pathogen and randomized data for the 7 other pathogens. We then compared the gene rankings obtained by PMM performed using both datasets for the selected pathogen. The results for *L. monocytogenes* are illustrated in Figure 3.7.A. The results for the other pathogens are given in Section 3.3.4. The high correlation shows that the addition of parallel screens had only a mild effect on the overall gene ranking. However, when considering the number of significant genes (FDR < 0.4), PMM mainly added genes to the list of significant genes (6 novel significant genes for *L. monocytogenes*) and only few genes (1 for *L. monocytogenes*) were dropped off the list. In general, we concluded that parallelism is adding novel significant genes. Almost no genes are lost through parallelism. The few lost hit genes had high FDR values, just slightly below the selected threshold FDR < 0.4.

In the next step we analyzed the differences between the resulting gene rankings of the tested methods. Differences in gene rankings between PMM and other hit ranking methods were not very strong (see Figure 3.7.B for MTT compared to PMM for *L. monocytogenes*). Genes that had a high sharedness score and had an effect on the screen of interest (in particular MTOR and TGFBRI/2 for *L. monocytogenes*) gained statistical power from the simultaneous analysis and were pushed up in the gene ranking. On the other hand, we observed that PMM detected several genes with low sharedness scores, indicating that unique hits were not neglected (see also Figure 3.10).

In order to evaluate the biological relevance of the observed hits, we calculated pathway enrichment scores separately for each pathogen
Figure 3.8: **GSEA pathway enrichment** The figure illustrates the GSEA results for the hit ranking methods PMM, MTT, and RSA. The numbers on top of the heatmap give the number of highly significant pathways (GSEA FDR < 0.2) for each hit detection method and pathogen. The significant pathways are highlighted with a red square.
by the Gene Set Enrichment Analysis (GSEA) algorithm (Subramanian et al. 2005) using as input the results from the three hit ranking algorithms PMM, MTT, and RSA. We selected all pathways that were significant (GSEA pathway enrichment FDR score < 0.2) for at least one pathogen and method pair and focused on hits that reduce infection levels (see Section 3.4.11 for more details). The results are illustrated in Figure 3.8. The heatmap illustrates the observed significant pathways. The number of enriched (GSEA FDR < 0.2) pathways for each method was an indication that PMM detected biologically more relevant hit genes than the other methods.

By assuming that most pathways in the GSEA database are biologically valid, we would expect that better hit detection methods give a higher number of enriched pathways than less powerful hit detection methods. However, we only screened kinases and the applicable pathways are limited to those that are highly enriched in phosphorylation events and it may be that some pathogens do not show strong enrichments within this set of pathways. Moreover, differences in pathway enrichments between methods may have occurred because they treated missing values differently. Therefore, the enrichment results should be evaluated with caution.

3.3 Additional Analysis Results

In this section we present additional analysis results which were not published as main results in Rämö, P and Drewek, A et al. 2014, but support our finding on the strength of PMM, as well as complete previously given examples by showing results for all the pathogens.

3.3.1 Cell Count Readout

We fitted the PMM also to the cell count readout. The fit yielded the same $c_{pg}$ scores for all the pathogens (all random effects $b_{pg}$ were
estimated to 0) and therefore, the model provided the same FDR values for all genes (see Figure 3.9). An explanation is that the cell count readout is a pathogen independent readout, whereas the infection index bases on the pathogen specific biological pathways leading to infection.

![Figure 3.9: Summary of PMM hits for the cell count readout. The heat map shows all genes which were significant (FDR < 0.4) at least for one pathogen. The colors correspond to the estimated c/pg values. NA refers to missing values, i.e. data was removed due to experimental bias.]

### 3.3.2 Similarities and Differences

We compared the gene ranking obtained by PMM with the gene rankings of MTT and RSA. In previous section we claimed that the differences in gene rankings were not very strong and showed as an example the results for *L. monocytogenes*. Figure 3.10 shows the comparisons between all method pairs for all the pathogens. The correlations between the three different methods are relatively high. However, there are several genes with high sharedness that were pushed up in the gene ranking by PMM.
Figure 3.10: Differences in rankings between the studied hit scoring methods. The y-axis shows the rank of a gene given by one method and the x-axis the rank defined by another method for a certain pathogen. The x and y labels indicate the methods by which the ranks were determined. The dot size corresponds to the sharedness score of each gene.
3.3.3 False Negatives

Both false positives and false negatives contribute to the low correlation between independent siRNA screens. False negatives usually arise from siRNAs that do not show a phenotype beyond the experimental background or siRNAs that do not perform a full knockdown given the chosen experimental conditions. Quality control in siRNA screening is certainly important to limit false negatives. Our plate layout includes several control wells to ensure that the experiments worked well. Moreover, we excluded from the analysis those genes that cause cytotoxicity. The determination of the number of false negatives is almost impossible without knowing the ground truth hits or making any further assumption. Using the same simulation set-

Figure 3.11: False Negatives. The figure shows the DET-curves for PMM, MTT and RSA applied on simulated data for three different scenarios (containing only hits that were shared between all pathogens, unique hits for all pathogens and mixed hit structure with unique and shared hits). The dashed and solid lines indicate whether the hits were generated by a low or high shift away from zero.
tings as for the calculation of the ROC curves (see Section 3.4.9 for details), we compared false negatives of the PMM and the reference method. The results show that the false negative rate of PMM is not higher than with the other commonly used methods (see Figure 3.11).

3.3.4 Parallelism of Screens

We analyzed by a resampling study how simultaneous modeling exploits the parallel structure of the data. Therefore, we compared the estimated $c_{pg}$ scores to re-estimated $c_{pg}$ scores originating from a re-fitted PMM based on data where we randomized the data of the other 7 parallel assays (see Section 3.4.10). The results were illustrated for *L. monocytogenes*. Figure 3.12 shows the results for all eight pathogens. Parallelism has only minor influences on the individual rankings. It mainly adds genes to the list of significant genes (red points), considering FDR < 0.4 for a gene to be significant. There is almost no loss (blue points).

3.3.5 Phylogenetic Relations

Despite the variety, we can classify groups of closely and distantly related pathogens among the studied bacterial pathogens within the InfectX consortium. All bacterial pathogens are gram negative bacteria, except for *L. monocytogenes*. The two bacteria *S. typhimurium* and *S. flexneri* are closely enterobacteria, whereas *B. henselae* and *B. abortus* are closely related pathogens belonging to the group of alphaproteobacteria. A comparison of sharedness between these groups did not reflect their phylogenetic relations. We found the pathogen pair *L. monocytogenes* and *B. abortus* at the top of the ranking with respect to sharedness. It was followed on second position by *B. abortus* and *S. typhimurium*. The two pathogens *S. typhimurium* and *S. flexneri* were almost at the end of the ranking. This finding agree with the notion that the entry mechanism of pathogens evolved relatively late in evolution and thus do not reflect phylogenetic relatedness. Therefore, the sharedness score does not introduce any bias.
Figure 3.12: **Power of Parallelism.** The y-axis shows the estimated $c_{pg}$ scores for the pathogen indicated in the title. The x-axis shows $c_{pg}$ scores originating from a refitted PMM based on data where we randomized the other 7 parallel assays. The colors correspond to significant genes (FDR < 0.4): green is a hit in both cases, red is a hit in the original data, and blue is a hit in the randomized data.
3.4 Methods

In this methods section we give details how the data were pre-processed and explain the statistical methods that were presented or used for simulations in the previous two sections.

3.4.1 Data Preprocessing and Normalization

Prediction of siRNA on-target genes

A target gene for a specific siRNA is defined as a gene, which exhibits perfect complementarity within its coding region to this siRNA. This is not necessarily a one-to-one relation, i.e. siRNAs can potentially have multiple target genes. In order to identify these target relations, siRNA sequences were mapped against genomic transcript sequences from RefSeq (release hg19, downloaded 17.07.2012) and ENSEMBL (release GRCh37.67, downloaded 20.07.2012) using BLAST version 2.2.27. The BLAST parameter word size was set to 7. Transcript matches shorter than the length of the siRNA sequence, as well as matches with gaps were removed. In the final step transcript IDs were translated to gene IDs and only a unique match of the target gene ID considering both genomic data sources were reported for each siRNA.

Z-Scoring

Individual plates of screens cannot always be handled identically in the wet-lab. There are several approaches in the literature to correct effects from single plates or plate batches (Birmingham et al. 2009). We chose non-control based data normalization methods since negative controls (MOCK and SCRAMBLED) sometimes show non-typical phenotypes (such as relatively high cell number) and good positive controls were not always available for all pathogens before primary screening. We used Z-Scoring to normalize variations between plates as:

\[ x_{\text{new}} = \frac{x_{\text{old}} - \mu}{\sigma}, \]
where $\mu$ is the mean of all siRNA well readouts in the plate, $\sigma$ is the standard deviation of all siRNA well readouts in the plate, $x_{\text{old}}$ is the raw well readout and $x_{\text{new}}$ is the normalized well readout.

The non-control based normalization assumes that all genes are randomly distributed among all plates and that there are relatively few positive phenotype genes in the whole screen. Our data is approximately Gaussian distributed and only a relatively small number of data points are outliers in each plate. Figure 3.13 shows the data for the Adenovirus Dharmacon pooled screen as example plate. The infection indices are nearly Gaussian distributed. Only a minority of data points in the plate are outliers. Since the other plates and pathogens showed similar pictures, the assumptions are fulfilled in a good approximation.

In order to obtain comparable screens, we applied a second Z-scoring to the plate-corrected data. The assays of each pathogen were separately again corrected by its mean and its standard deviation to correct for laboratory effects since the experiments of each pathogen was performed in different laboratory.

Figure 3.13: Assumptions for Z-Scoring. The histogram and the QQ-plot visualize the readout data (non-normalized infection index) from a plate of the Adenovirus Dharmacon Pooled screen. The plots show a Gaussian distribution with slightly fatter tails.
Dependence on the Population Context

Infection phenotypes can depend on the population context, such as the total cell number (Snijder et al. 2012, Knapp et al. 2011, König et al. 2007). Figure 3.14 shows the infection index versus cell count for all pathogens. There is a slight dependency on the cell number for \( B. \) \( henselae \), \( L. \) \( monocytogenes \) and Vacinii virus). We applied non-parametric regression correction with the Lowess-method (Yang et al. 2002) in order to reduce the bias caused by the dependence. We used a sliding window of size 200 to go through the Z-scored, ranked cell number readouts. For each window we calculated the mean \( m \) and standard deviation \( s \). With those values we then calculated a Lowess-normalized value:

\[
x_{\text{new}} = \frac{x_{\text{old}} - m}{s}.
\]

The normalization was separately performed for the infection index of each pathogen and siRNA library.

3.4.2 Access to Data

The raw and normalized data sets, as well as the original microscope images of the kinome-wide InfectX screen are available on the InfectX openBIS data publication portal. The portal is located at:

URL: http://www.infectx.ch/dataaccess/
Username: rdgr2014
Password: IXPubReview

The pre-processed and normalized data is also available as example data in the PMM R-Package (Drewek 2015).
Figure 3.14: Dependence on the Population Context. The scatter plots show the dependencies of Z-Scored infection index with the cell number. The red lines correspond to the smoothed Lowess average estimate and the green lines to the +/- standard deviation. The underlying data is taken from the DHARMaCON pooled siRNA libraries.
3.4.3 Parallel Mixed Model (PMM)

PMM is composed of a linear mixed model and an assessment of the local False Discovery Rate (FDR).

**Linear Mixed Model**

We denote the readout of siRNA $s$ silencing gene $g$ for a pathogen $g$ as $y_{pgs}$. The linear mixed model of PMM is defined as the following linear model

\[
y_{pgs} = \mu_p + a_g + b_{pg} + \varepsilon_{pgs},
\]

\[
a_g \sim (0, \sigma_a^2), \quad b_{pg} \sim (0, \sigma_b^2), \quad \varepsilon_{pgs} \sim (0, \sigma_\varepsilon^2),
\]

where $\mu_p$ is the fixed effect for pathogen $p$ (typically close to 0 because of data Z-Scoring), $a_g$ is the gene effect overall pathogens, $b_{pg}$ is the gene effect within a pathogen and $\varepsilon_{pgs}$ denotes the error term. The sum of two random effects $a_g$ and $b_{pg}$ describes the total effect of the siRNAs within pathogen $p$. We define the estimated effect $c_{pg}$ for gene $g$ within pathogen $p$ as

\[
c_{pg} = a_g + b_{pg}.
\]

The parameters of the linear mixed model are estimated by maximizing the restricted maximum likelihood using the Newton-Raphson algorithm (Pinheiro and Bates 2000). We used the implemented version in the `lmer` function from the `lme4` R-package (Bates et al. 2015). This implementation allows also the use of weights, which are incorporated by a weighted maximum likelihood formulation. The weights are constant values where each constant corresponds to exactly one data point. For the InfectX data, each weight is associated with a single readout of an independent siRNA. The size of the weight indicates the precision of the information contained in the associated readout.

The linear mixed model assumes that the model residuals $\varepsilon_{pgs}$, as well as the random coefficients $a_g$ and $b_{pg}$ are normally distributed.
Figure 3.15: Assumptions of PMM. (a) The QQ-plot compares the observed quantiles of residuals from the PMM to the theoretical quantiles of a normal distribution. (b) The scatter plot shows the residuals of all combinations of genes and pathogens. (c) This scatter plot is a “Zoomed-in-version” of the scatter plot in (b) for both ends. The plot shows a random scatter around 0, indicating that there are no systematic errors in the estimation of PMM. (d) The QQ-plot compares the observed quantiles of the gene random effects ($a_g$) to the theoretical quantiles of a normal distribution. (e) The QQ-plot shows the gene random effect within a pathogen ($b_{pg}$).

Figure 3.15 shows diagnostic plots for the fitted linear mixed model. The QQ-plot shows that the normal distribution for residuals is only approximately satisfied (see Figure 3.15.A). There are outliers in the residuals with respect to positive and negative infection indices (marked with red points). To check whether the estimation of the linear mixed model is affected by these outliers, we refitted the linear mixed model without the red marked points. The resulting hit lists of the linear mixed model with and without the red points are almost identical. Moreover, the residuals are randomly distributed around zero for each gene within a pathogen (see Figure 3.15.B-C). This indicates that there is no systematic error in the estimation of the linear mixed model. Therefore, the linear mixed model fit is reli-
able concerning the residuals assumptions. The two other QQ-plots confirm the assumption of normal distribution for the random coefficients (see Figure 3.15.D-E).

Local False Discovery Rate

The observed distribution of the estimated $c_{pg}$ is a mixture of the null $f_0$ and the non-null distribution $f_1$. The null distribution describes the distribution of all genes that are no-hits. The non-null distribution corresponds to the genes that are hits, having either a positive or negative effect. The two distributions are assumed to differ only in the mean. The non-null distribution is shifted by a factor $\theta$ away from zero. We define the local false discovery rate as

$$fdr(c) = P(\text{No Hit} \mid c) = \frac{\pi_0 f_0(c)}{f(c)}$$

$$f_0 \sim (0, \sigma_a^2 + \sigma_b^2), \quad f_1 \sim (\theta, \sigma_a^2 + \sigma_b^2)$$

$$f(c) = \pi_0 f_0(c) + \pi_1 f_1(c)$$

where $\pi_0 = \text{proportion of true hits}$ and $\pi_1 = 1 - \pi_0$ according to Efron 2010. Following the estimation procedure of Efron et al. 2011, we separately estimate the three quantities needed for the estimation of $fdr(c)$. First we define a subset $A_0$ for which holds $f_1(c) = 0$. An estimate for the proportion of true hits is then obtained by

$$\hat{\pi}_0 = \frac{M(A_0)}{N \cdot \hat{f}_0(A_0)}$$

where $N := \text{number of genes}$ and $M(A_0) := \text{observed number of } c_i \text{ in } A_0$. An estimation for the null distribution $f_0(z)$ can be deduced by Maximum Likelihood estimation using a mixture of the binomial and normal distribution with respect to the subset $A_0$. As last, the estimate for the mixture distribution $f(z)$ can be obtained by a Poisson regression based on Lindsey’s method. We refer to Efron 2010 for more details.
3.4 Methods

Implementation of PMM

The Parallel Mixed Model (PMM) approach is not only suitable for the hit selection and cross-comparison of the InfectX RNAi screens, it can be applied to all kind of RNAi experiments that are performed in parallel under several conditions. For example, we could think of the measurements or readouts from cells under RNAi knock-down, which are infected with several other pathogens or which are grown from different cell lines. We, therefore, built an R-package of the same name as the method PMM and submitted it to Bioconductor (Drewek 2015). All related documentation is available on the Bioconductor website.

3.4.4 Moderated T-Test

As a reference method, we used Moderated T-Test (MTT) (Smyth 2004). It tests whether the observed distribution of a sample has a mean equal to 0. The difference to the one-sample t-test is that the test statistic includes prior information on the different variances of the siRNAs within the genes. It therefore assumes that the standard deviation of the test samples are chi-squared distributed. We performed MTT using the R implementation presented in Smyth 2005. Our data satisfies the method assumptions (see Figure 3.16.A). The QQ-Plot of the observed quantiles of the sample standard deviation versus a Chi-Squared distribution shows a straight line for all eight pathogens.

Due to multiple testing errors, the p-values of MTT cannot be directly used in large-scale screening (Storey 2002, Storey and Tibshirani 2003, Prummer 2012, Benjamini and Hochberg 1995). We used the Storey’s multiple testing correction (Storey and Tibshirani 2003). This approach assumes that the p-values have a flat distribution with a possible peak at the low end. The histograms illustrates that our p-values fulfill this assumption for all pathogens (see Figure 3.16.B).
3.4.5 Redundant SiRNA Analysis

The Redundant siRNA Analysis (RSA) ranks all siRNAs targeting a given gene over all siRNAs in the screens. It assigns the p-values for each gene based on a hypergeometric distribution that indicates whether the distribution of ranks of this gene is shifted significantly towards low ranks (König et al. 2007). RSA was run using the R-package RSA (König et al. 2007) with parameters: $l = -1.5$ and $u = 1$, where $l$ refers to the threshold where a single siRNA readout is considered to be true positive at the low end and $u$ refers to the threshold where a single siRNA readout is considered to be true positive at the high end.

3.4.6 Power Analysis of PMM

We showed by a resampling study that the simultaneous analysis of parallel screening data gains statistical power in finding shared hits. The results are shown in Figure 3.3. The resampling study was performed as follows. We chose a gene $g$ and a pathogen $p$ for which
we wanted to show the increase in power by simultaneously using the parallel screening structure. We resampled data for a fixed number of siRNAs ($ns = 2, \ldots, 11$) and a fixed number of pathogens ($np = 2, \ldots, 8$) from the full dataset. Therefore, we chose randomly ($np - 1$) pathogens and added additionally pathogen $p$ and we sampled in a next step $ns$ siRNA sets from the full available set of siRNAs for every gene within all sampled pathogens. We applied PMM on the sampled data and we reported the rank of gene $g$ within pathogen $p$. All the steps were repeated 1000 times for each combination of $ns$ and $np$. As a last step we calculated for each combination the mean and variance of the rank for gene $g$ within pathogen $p$. For the case of $np = 1$, we applied the same resampling procedure using MTT.

Our choice for $p$ and $g$ was based on a-priori information on hit structure. We took a unique hit ($g$: MET, $p$: L. monocytogenes), a shared hit ($g$: MTOR, $p$: Vaccinia virus) and a non-hit ($g$: ALK, $p$: B. abortus). For the resampling we omitted genes that have less than 6 siRNA sets, in order to have a good resampling basis.

### 3.4.7 Test for Stability

We analyzed the stability of the obtained gene rankings of PMM and the reference methods MTT and RSA by a resampling study. We resampled with replacement 1000 datasets from the full data, taking for each gene the same number of siRNA replicates as in the full dataset. The three hit ranking methods PMM, MTT and RSA were then applied to each resampled dataset and the corresponding three rankings were saved. We ranked the results of PMM according to the absolute value of the estimated $c_{pq}$ effects. The results of the reference methods were ranked for MTT by the absolute values of the estimated mean and for RSA by the $\log(p)$ values. We took absolute values to take simultaneously into account down and up hits. From the 1000 rankings we calculated the number of genes that appear with high probability ($\text{prob} > 0.9$ and $\text{prob} > 0.7$) in the top $k$ ($k = 1, \ldots, 50$) of the ranking.
3.4.8 Hit Reproducibility

We studied the reproducibility of the top ranked genes by a cross-validation analysis. For this we only used data coming from the siRNAs of all unpooled libraries (11 in total). In each run, we ran PMM, MTT, and RSA on a subset of the data consisting of 10 individual siRNAs and used the remaining siRNA set as test set. We ranked the results of PMM according to the absolute value of the estimated \( c_{pg} \) effects. The results of the reference methods were ranked for MTT by the absolute values of the estimated mean and for RSA by the \( \log(p) \) values. The genes of the test set were ordered by the absolute value of infection score. We counted the number of genes that appeared in top \( k \) \((k = 1, \ldots, 100)\) in both the training and test sets. We determined the counts separately for each pathogen and averaged them in the end.

3.4.9 Goodness of Gene Rankings

We performed a data simulation with a-priori known hit structure in order to evaluate the false positives in our ranking. We applied all three hit-ranking methods to the simulated data. We simulated data by generating 1000 normally distributed screens (mean = 0, std = 0.5) for eight pathogens, taking 4 siRNAs each. Hits were incorporated in the simulated screens by randomly selecting about 10% of the genes (80 out of 826) and shifting them away from zero. The shift was determined by a uniformly distributed random variable. We used the interval \([a, b]\) as parameter for the uniform distribution. We simulated two cases: a low shift with \(a = 0.2\) and \(b = 0.3\) and a big shift with \(a = 0.4\) and \(b = 0.5\). In addition, we simulated three cases of possible hit structures: In the first case the hits were different for each pathogen (80 unique hits per pathogen), in the second case all hits were shared between the pathogens (same 80 hits for all pathogens) and in the third case we generated mixed hits (20 unique hits, 20 hits shared between two pathogens, 20 hits shared between four pathogens and 20 hits shared between all eight pathogens).
PMM, MTT, and RSA were applied to the simulated data and the ranking was saved. For PMM the results were ranked according to the absolute value of the estimated $c_{pg}$ effects, for MTT the ranking was done with respect to the absolute values of the estimated mean and for RSA the ranking based on the $\log(p)$ values. For every ranking list we counted the number of true positives, true negatives, false positives and false negatives in the top $k$ ($k = 1, \ldots, 826$) and computed the true positive rate ($\text{TPR} = TP/(TP + FN)$) and the false positive rate ($\text{FPR} = FP/(FP + TN)$).

### 3.4.10 Influence of Parallelism

We analyzed how simultaneous modeling exploits the parallel structure of the data by a small resampling study. For one selected pathogen $p$ we generated 1000 new datasets by fixing the data of $p$ and randomizing the data of the other 7 pathogens. We applied PMM to each dataset and saved the resulting ranking of pathogen $p$. In the next step we aggregated the 1000 rankings by taking the average over the $c_{pg}$ scores. We compared the averaged scores to the gene rankings obtained by PMM with the original dataset. We independently performed the resampling study for each pathogen.

### 3.4.11 Gene Set Enrichment Analysis

We evaluated the biological relevance of the hits determined by PMM, MTT and RSA by a gene set enrichment analysis. We calculated the pathway enrichment scores separately for each pathogen by the Gene Set Enrichment Analysis (GSEA) algorithm. We used the results from the three hit ranking algorithms PMM, MTT, and RSA as input. The Gene Set Enrichment Analysis (GSEA) was run using the Java-package “gsea2-2.1.0.jar” and the curated canonical pathways “c2.cp.v4.0.entrez.gmt” (Subramanian et al. 2005). The following settings were used for the parameters: collapse was set to false, mode to Max_probe, norm to meandiv, nperm to 1000, scoring_scheme to classic, include_only_symbols to true, make_sets to true, set_max to
500 and set $\min$ to 7. We limited ourselves to pathways that had at least 7 kinases within the pathway in order to avoid bias towards too small pathways. We decided to use the “classic” approach in order to keep different hit scoring methods (PMM, MTT, and RSA) comparable.

### 3.5 Biological Inquiry on Detected Significant Genes

The performed screens yield several interesting hits of which most are novel to the corresponding pathogen (see Figure 3.5.A). Many of the strongest hits, including $MTOR$, $TGFBR1$ and $TGFBR2$ for negative hits and $ILK$ for positive hits, were shared between most of the studied pathogens. This was also illustrated by the sharedness scores of detected hit genes. Many of the strongest shared hits were related to $SRC$, $MTOR$, or $CDK$ pathways. Although $SRC$ and $CDK4$ were not part of the hit lists ($q_{pg} < 0.4$) for any of the pathogens, they exhibited consistent semi-strong effect for most pathogens. A network analysis of the hit genes showed that several of the shared hits can be described as “network hubs”. They are involved in many cellular processes and highly connected to other genes (including $MTOR$ and $SRC$) (see Figure 3.5.B) (Szklarczyk et al. 2011).

$MTOR$ is a mammalian target of rapamycin, serine/threonine protein kinase that regulates cell growth, cell proliferation, cell motility, cell survival, protein synthesis, and transcription. The involvement of $MTOR$ in $Adenovirus$, $Polioivirus$, $Enterovirus71$, $Coxackievirus$, $Vaccinia virus$ and other pathogens has already been established (Mercer et al. 2012, Snijder et al. 2012, Sivan et al. 2013). Our data also reproduced the established role of $MTOR$ during $S. typhimurium$ infections, since $S. typhimurium$ depends on a reactivation of $MTOR$ during its course of infection in order to escape autophagy (Tattoli et al. 2012).
Interestingly, *TGFBR1* and *TGFBR2* came up both as strong hits for many pathogens. *TGFBR1* and *TGFBR2* proteins must heterodimerize to form a functional *TGF-β* receptor at the plasma membrane. Their similar strong infection reducing knock-down phenotypes, seen in most independent pathogen screens, indicated the validity of these hits and suggested a broad, yet poorly understood, function of this membrane protein for various pathogens. In particular, there are suggestions (Elfakia and Al-Hokailb 2009) that the *TGFB* pathway might be important for *B. abortus* infection. Patients with chronic brucellosis have an increased *TGF-β* production and this could aid infection by depressing lymphocyte functions.

Despite the overall similarity of infection patterns between pathogens, most pathogens also contained hits that were specific for the pathogen (for example *MET* for *L. monocytogenes*, *NTPCR* for *B. henselae*, and *ETNK1* and *ULK1* for *Rhinovirus*). Some of the hit genes have previously been found to be effectors, for example *MET* for *L. monocytogenes*. *L. monocytogenes* enters host cells by triggering signaling cascades activated through interaction of bacterial internalin A (InlA) or InlB with the adherens junction protein E-cadherin or the hepatocyte growth factor receptor *MET* (Pizarro-Cerdá et al. 2012) respectively. Since E-cadherin is not expressed in HeLa cells, which were used for our siRNA screens, the *INLB*/*MET* pathway is the only route of entry in the cellular system. In fact, *MET* (Shen et al. 2000) was one of the strongest hits for *L. monocytogenes*.

The exact roles of most hit genes of all pathogens are largely unknown, but several hit genes create interesting hypotheses for follow-up. For example, it was proposed based on microRNA analysis of infected macrophages, that *AMPK* might be a target gene that promotes intracellular survival during *B. abortus* infection (Zheng et al. 2012). *PIK3R3* (p55-gamma; Phosphatidylinositol 3-Kinase 55 KD a Regulatory Subunit Gamma) a semi-strong hit for several pathogens in our data was identified as a hit in an RNAi screen of drosophila S2 cells, in agreement with the importance of *PI3K* during *B. abortus* infection (Qin et al. 2008). *PIK3CA* probably plays a role also in *B. henselae* infection through actin modulation. *PIK3CA* levels influ-
ence $RHOA$ and $RAC1$, which are involved in actin dynamics (Cain et al. 2010). Furthermore, $PIK3CA$ is involved in $\textit{PIP3}$ production, which is a signaling molecule and has recently been shown to be related to the formation of dynamic F-actin-related structures (Kakumoto and Nakata 2013). $ULK1$ (unc-51 like autophagy activating kinase 1) plays an important role in autophagy as well as \textit{Hepatitis C virus} infection. Therefore, $ULK1$ has a possible link to \textit{Rhinovirus} induced autophagy. $COL4A3BP$ is possibly linked to \textit{Rhinovirus} entry through ceramide-enriched membrane platforms (Grassm´e et al. 2005) since $COL4A3BP$ specifically phosphorylates the N-terminal region of the non-collagenous domain of the alpha 3 chain of type IV collagen, known as the Goodpasture antigen, also involved in ceramide intracellular transport (from ER to Golgi).

### 3.6 Discussion and Conclusions

We produced a uniquely kinome-wide high-content siRNA dataset, in terms of used siRNA libraries (11 single siRNAs and one pool) for eight different pathogens. Our highly unified protocols and common image analysis as well as similar data analysis pipelines enabled a direct comparison between the phenotypic readouts of the different pathogen screens. The unified structures of the datasets allowed the discovery of shared mechanisms between the studied pathogens.

Using our novel statistical approach PMM we detected several interesting and new hits from our kinome-wide pathogen screens. The hits will require further follow-up work in order to understand the exact biological mechanisms of the genes. In addition, we discovered shared effector genes including $MTOR$, $TGFBR1$ and $TGFBR2$ that were strong hits for almost all studied pathogens. In particular, the obtained sharedness scores indicated whether a hit gene has a very specific function for a single pathogen or a more generic cellular function that is shared between many pathogens. Thus it gave us the first indications of the gene’s roles. Pharmaceutically oriented follow-up studies could take advantage of this concept. For example, if we were
interested in general regulators we could focus on genes with high sharedness scores. On the other hand, regulators that have a very specific effect and a low sharedness score could probably have fewer side effects.

We showed that the reliability of hit scoring in individual RNAi screens improved by using PMM that takes advantages of the parallelism in RNAi screening. PMM can, in principle, be applied to any kind of parallel RNAi screens almost independently of the underlying biology or field of application as long as the readouts of the screens are measured on the same scale. We can often obtain this by applying Z-Scoring or similar normalization methods to the well readouts. The difference to other approaches aiming at the comparison of independent parallel RNAi screens is that PMM takes simultaneously all screening data into account. For example, the hit lists were derived by separate statistics on each screen for the comparison of insect and human data in Sessions et al. 2009. Taking all data into the analysis can increase the statistical power. Based on our results, we expect that the more similar the parallel screens are in the sense of biological focus or protocols, the more statistical power can be gained from the simultaneous analysis. Even a slight overlap between the underlying biological pathways of the parallel screens can improve the hit detection in individual screens without compromising the detection of unique hits for any individual screens.

Assuming that the large-scale RNAi screening community reaches standardized data publication and shares these standards through projects such as MIARE and GenomeRNAi, the PMM approach could combine a vast number of different RNAi screens performed in different laboratories worldwide that used the same siRNA libraries. In principle and as a vision, this opens up great opportunities for simultaneous statistical approaches such as PMM. Every new screen could potentially gain statistical power by using the public resources. In addition, the PMM can potentially be used to gain power for secondary validation screens. Such validation screens are typically performed with several independent siRNAs targeting the same gene under various conditions and PMM is here directly applicable.
A beneficial feature of the PMM is the possibility to assign weights to the siRNAs. The weights can incorporate a-priori information about the performance of individual siRNAs and their phenotypic readout. This concept of weighting can be expanded over what we presented in Section 3.2.4. In particular, statistical and bioinformatics analyses on seed sequence induced off-target effects could potentially be used as basis for weights. Naturally any additional high-throughput data, such as proteomics analyses on cells under siRNA perturbations, or genomic analyses on specific cell lines, could be used to assign realistic siRNA weights to improve the hit scoring.

We aimed to take a step forward in determining minimal requirements for image-based RNAi screening data publication. All the raw images, library metadata, individual cell measurements, and well measurements are publicly available through our openBIS based publication portal. In addition, we provide easy-to-access data aggregates in standardized tabular formats with all the necessary metadata information. Our uniquely wide datasets provide a large resource for infection biologists, image analysts, and statisticians for future research.
Appendix Chapter 3

3.A Biological Protocols

3.A.1 Assays Performed

The InfectX consortium screened all studied pathogens with four different siRNA libraries (Dharmacon pooled, Dharmacon unpooled, Ambion unpooled, and Qiagen unpooled). Table 3.1 reports the number of replicates of the different pathogen assays. The number varies since assays needed to be removed due to different problems (e.g. transfection did not work or an older protocol was used that gave a weak fluorescent signal). We used the Dharmacon pooled library for assay optimization. Therefore there the number of replicates varies most. In all analyses we averaged data from replicates in order to have only one value per library and gene.

<table>
<thead>
<tr>
<th>Pathogen</th>
<th>Dharmacon 1 pooled siRNA</th>
<th>Ambion 3 unpooled siRNAs</th>
<th>Qiagen 4 unpooled siRNAs</th>
<th>Dharmacon 4 unpooled siRNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenovirus</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B. abortus</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B. henselae</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>L. monocytogenes</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Rhinovirus</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S. flexneri</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S. typhimurium</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Vaccinia virus</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.1: Replicates performed for different pathogens and libraries.

3.A.2 Main Differences in Protocols

We designed specific protocols for each pathogen so that the resulting phenotypes (infection index and cell number) ended up as similar as possible between all the screens. Table 3.2 summarizes the main differences in protocols.
Table 3.2: Table of pathogens and main differences in protocols.

<table>
<thead>
<tr>
<th>Pathogen</th>
<th>Actin Stain</th>
<th>DNA Stain</th>
<th>Pathogen entry time</th>
<th>Multiplicity of infection (MOI)</th>
<th>Seeded cell number per well</th>
<th>Venetian virus</th>
<th>Vaccinia Virus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenovirus</td>
<td>CFP</td>
<td>CFP</td>
<td>16h</td>
<td>0.1</td>
<td>700</td>
<td>Adenovirus AD2-E3B-eGFP</td>
<td>CFp: DY-647-phalloidin</td>
</tr>
<tr>
<td>B. abortus</td>
<td>CFP</td>
<td>CFP</td>
<td>30h</td>
<td>400</td>
<td>300</td>
<td>B. abortus B. henselae ∆bepG</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>B. henselae 2308 pJC43 (aphT::GFP)</td>
<td>CFP</td>
<td>CFP</td>
<td>4h</td>
<td>10000</td>
<td>0.01</td>
<td>B. henselae</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>L. monocytogenes</td>
<td>CFP</td>
<td>CFP</td>
<td>20min</td>
<td>10000</td>
<td>0.01</td>
<td>L. monocytogenes</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>R. typhi</td>
<td>CFP</td>
<td>CFP</td>
<td>30min</td>
<td>200</td>
<td>0.125</td>
<td>R. typhi</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. flexneri M90T ∆virG pCK100</td>
<td>CFP</td>
<td>CFP</td>
<td>30min</td>
<td>15</td>
<td>600</td>
<td>S. flexneri</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. typhimurium (1) S. typhimurium (2) sopE2sipAsopB (S. TmSopE pM975)</td>
<td>CFP</td>
<td>CFP</td>
<td>20min</td>
<td>80</td>
<td>800</td>
<td>S. typhimurium</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. enterica 907T ΔvirC pCk100</td>
<td>CFP</td>
<td>CFP</td>
<td>15</td>
<td>600</td>
<td>600</td>
<td>S. enterica</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. flexneri M90T ∆virG pCK100</td>
<td>CFP</td>
<td>CFP</td>
<td>30min</td>
<td>15</td>
<td>600</td>
<td>S. flexneri</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. flexneri M90T ∆virG pCK100</td>
<td>CFP</td>
<td>CFP</td>
<td>30min</td>
<td>15</td>
<td>600</td>
<td>S. flexneri</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. enterica 907T ΔvirC pCk100</td>
<td>CFP</td>
<td>CFP</td>
<td>15</td>
<td>600</td>
<td>600</td>
<td>S. enterica</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. enterica 907T ΔvirC pCk100</td>
<td>CFP</td>
<td>CFP</td>
<td>15</td>
<td>600</td>
<td>600</td>
<td>S. enterica</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. enterica 907T ΔvirC pCk100</td>
<td>CFP</td>
<td>CFP</td>
<td>15</td>
<td>600</td>
<td>600</td>
<td>S. enterica</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. enterica 907T ΔvirC pCk100</td>
<td>CFP</td>
<td>CFP</td>
<td>15</td>
<td>600</td>
<td>600</td>
<td>S. enterica</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. enterica 907T ΔvirC pCk100</td>
<td>CFP</td>
<td>CFP</td>
<td>15</td>
<td>600</td>
<td>600</td>
<td>S. enterica</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. enterica 907T ΔvirC pCk100</td>
<td>CFP</td>
<td>CFP</td>
<td>15</td>
<td>600</td>
<td>600</td>
<td>S. enterica</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. enterica 907T ΔvirC pCk100</td>
<td>CFP</td>
<td>CFP</td>
<td>15</td>
<td>600</td>
<td>600</td>
<td>S. enterica</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. enterica 907T ΔvirC pCk100</td>
<td>CFP</td>
<td>CFP</td>
<td>15</td>
<td>600</td>
<td>600</td>
<td>S. enterica</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. enterica 907T ΔvirC pCk100</td>
<td>CFP</td>
<td>CFP</td>
<td>15</td>
<td>600</td>
<td>600</td>
<td>S. enterica</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. enterica 907T ΔvirC pCk100</td>
<td>CFP</td>
<td>CFP</td>
<td>15</td>
<td>600</td>
<td>600</td>
<td>S. enterica</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. enterica 907T ΔvirC pCk100</td>
<td>CFP</td>
<td>CFP</td>
<td>15</td>
<td>600</td>
<td>600</td>
<td>S. enterica</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. enterica 907T ΔvirC pCk100</td>
<td>CFP</td>
<td>CFP</td>
<td>15</td>
<td>600</td>
<td>600</td>
<td>S. enterica</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
<tr>
<td>S. enterica 907T ΔvirC pCk100</td>
<td>CFP</td>
<td>CFP</td>
<td>15</td>
<td>600</td>
<td>600</td>
<td>S. enterica</td>
<td>CFP: DY-647-phalloidin</td>
</tr>
</tbody>
</table>
| S. enterica 907T ΔvirC pCk100 | CFP | CFP | 15 | 600 | 600 | S. enterica | CFP: DY-647-phalloidin | RFP: DY-547-phalo
3.A Biological Protocols

3.A.3 Wet-lab Protocols

General Protocol

The InfectX consortium conducted all experiments in a 384-well plate format. HeLa CCL-2 (ATCC) cells were maintained at 37°C and 5% CO\textsubscript{2} in Dulbecco Modified Eagle Medium (DMEM, Invitrogen) supplemented with 10% inactivated FCS (Invitrogen). We achieved the RNA interference directed against human kinases and kinase-associated genes (826 genes in total) using commercially available siRNA libraries. In addition, all plates contained general siRNA controls for transfection efficiency and toxicity (e.g. Kif11), as well as, control siRNAs for infection effects of each pathogen assayed. Next, 25\( \mu \)l of RNAiMAX/DMEM (0.1\( \mu \)l/24.9\( \mu \)l) mixture was added to each well of the screening plates containing 1.6pmol siRNA diluted in 5\( \mu \)l RNase-free ddH\textsubscript{2}O. Screening plates were thereafter incubated at room temperature (RT) for 1h. After the incubation we added a pathogen assay-specific number of HeLa CCL-2 cells (see Table 3.1) per well in a volume of 50\( \mu \)l DMEM/16% FCS. This resulted in a final FCS concentration of 10% (only Adenovirus screens contained 6.7% final FCS). Plates were incubated at 37°C and 5%CO\textsubscript{2} for 72h. Then cells were infected with a pathogen (see below for the pathogen-specific protocols). Afterwards the cells were fixed using paraformaldehyde (PFA) and stained for DNA, F-actin and infection specific markers. Screening plates were sealed prior to microscope imaging.

Adenovirus Protocol

All liquid handling stages of infection, fixation, and immunofluorescence staining were performed on the automated pipetting system Well Mate (Thermo Scientific Matrix) and the washer Hydrospeed (Tecan). Ad2\textsubscript{-ΔE3B-eGFP} (Adenovirus) was added to the cells at a multiplicity of infection (moi) of 0.1 in 10\( \mu \)l of an infection media/FBS (DMEM supplemented with L-glutamine, 10% FBS, 1% Pen/Strep, Invitrogen). Screening plates were incubated at 37 °C for 16h.
The cells were fixed by adding 21\(\mu l\) of 16\% PFA directly to the cells in culture media for 45 min at room temperature or long-term storage at 4\(^\circ\)C. Cells were washed twice with PBS/25 mM NH\(_4\)Cl, permeabilized with 25 \(\mu l\) 0.1\% Triton X-100 (Pharmaciebiothek). After 2 washes with PBS the samples were incubated at RT for 1h with 25\(\mu l\) staining solution (PBS) containing DAPI (1\(\mu g/ml\), Sigma-Aldrich) and DY-647-phalloidin (U/ml, Dyomics), washed 2 times with PBS and stored until imaging in 50\(\mu l\) PBS/Na\(_3\)N.

**Bartonella henselae Protocol**

*Bartonella henselae* (ATCC49882T \(\Delta\)bepG) were grown on Columbia base agar (CBA) plates supplemented with 5\% defibrinated sheep blood (Oxoid) and 50\(\mu g/ml\) kanamycin. The bacteria contained plasmid pCD353 (Dehio et al. 1998) for IPTG-inducible expression of GFP. The bacteria were incubated at 35\(^\circ\)C in 5\%CO\(_2\) for 72h before re-streaking them on fresh CBA and further growth for 48h. Cells were washed once with M199 (Invitrogen)/10\% FCS using a plate washer (ELx50-16, BioTek). Afterwards cells were infected with *B. henselae* at a MOI of 400 in 50\(\mu l\) of M199/10\% FCS and 0.5mM IPTG (Applichem) and were incubated at 35\(^\circ\)C in 5\%CO\(_2\) for 30h.

Fixation at room temperature was performed using a Multidrop 384 (Thermo Scientific). Cells were washed with 50\(\mu l\) PBS, fixed in 20\(\mu l\) of 3.7\% PFA for 10min, and washed once more with 50\(\mu l\) PBS. The staining was performed on a Biomek liquid handling platform. The fixed cells were washed twice with 25\(\mu l\) of PBS and blocked in PBS/0.2\% BSA for 10min. Extracellular bacteria were labeled with a rabbit serum 2037 against *B. henselae* (Dehio et al. 1997) and a secondary antibody goat anti rabbit Alexa Fluor 647 (Jackson Immuno) in PBS/0.2\% BSA. These antibodies were incubated for 30min each. Both incubations were followed by two washings with 25\(\mu l\) of PBS. Cells were then permeabilized with 20\(\mu l\) of 0.1\% Triton X-100 (Sigma) for 10min. Cells were then washed twice with 25\(\mu l\) PBS, followed by an addition of 20\(\mu l\) of staining solution (PBS containing 1.5U/ml DY-547-Phalloidin (Dyomics) and 1ug/ml DAPI (Roche)). The cells were incubated 30min in the staining solution. After incu-
bation the cells were washed twice with 25µl PBS, followed by the addition of 50µl PBS.

**Brucella abortus Protocol**

*Brucella abortus* 2308 pJC43 (aphT::GFP) (Celli et al. 2005) were grown in tryptic soy broth (TSB) medium that contained 50µg/ml kanamycin for 20h at 37°C and was shaked (100 rpm) to an OD of 0.8-1.1. Per well 50µl of DMEM/10% containing bacteria was added to obtain a final MOI of 10000 using a cell plate washer (ELx50-16, BioTek). Plates were then centrifugated at 400g for 20min at 4°C to synchronize bacterial entry. After 4h incubation at 37°C and 5%CO₂, extracellular bacteria were killed by exchanging the infection medium by 50µl medium supplemented with 10% FCS and 100µg/ml gentamicin (Sigma).

Cells were fixed with 3.7% PFA for 20min at RT with the cell plate washer (ELx50-16, BioTek). Staining was performed using a Biomek liquid handling platform. Cells were washed twice with PBS and permeabilized with 0.1% Triton X (Sigma) for 10min. Then, cells were washed twice with PBS, followed by addition of 20µl of staining solution which includes DAPI (1µg/ml, Roche) and DY-547-phalloidin (1.5 U/ml, Dyomics) in 0.5% BSA in PBS. Cells were incubated with staining solution for 30min at RT and were washed twice with PBS, followed by final addition of 50µl PBS.

**Listeria monocytogenes Protocol**

*L. monocytogenes* EGDe.PrfA*GFP were washed three times with PBS and then diluted in DMEM supplemented with 1% FCS. Cells were infected at a MOI of 25 in 30µl infection medium per well. After centrifugation at 1000rpm for 5min and incubation for 1h at 37°C in 5%CO₂ all extracellular bacteria were killed by exchanging the infection medium by 30µl DMEM supplemented with 10% FCS and 40µg/ml gentamicin (Gibco). Both medium exchange steps were carried out with a plate washer (ELx50-16, BioTek).
After additional 4h at 37°C in a 5% CO₂ atmosphere, cells were fixed for 15min at RT by adding 30µl of 8% PFA in PBS to each well using a multidrop 384 device (Thermo Electron Corporation). PFA was removed by four washes with 500µl PBS per well using the Power Washer 384 (Tecan). Fixed cells were stained for nuclei, actin and bacterially secreted InlC. First, cells were incubated for 30min with 10 µl/well of primary staining solution (0.2% saponin, PBS) containing rabbit derived anti-InlC serum (1:250). After four washes with 40 µl PBS per well cells were stained with 10µl/well of the secondary staining solution (0.2% saponin, PBS) containing Alexa Fluor-546 coupled anti-rabbit antibody (1:250, Invitrogen), DAPI (0.7 µg/ml, Roche), and DY-647-Phalloidin (2U/ml, Dyomics). After four washes with 40µl PBS per well, the cells were kept in 40µl PBS per well. The staining procedure was carried out with a Tecan freedom evo robot.

**Rhinovirus Protocol**

All liquid handling stages of infection, fixation, and immunofluorescence staining were performed on the automated pipetting system Well Mate (Thermo Scientific Matrix) and washer Hydrospeed (Tecan). For infection assays with human *Rhinovirus* serotype 1a (HRV1a) were carried out as described in Jurgeit et al. 2010, except that the anti-VP2 antibody Mab 16/7 was used for staining of the infected cells as described earlier (Jurgeit et al. 2012, Mosser et al. 2002). *Rhinovirus* at a MOI of 8 was added to cells in 20 µl of an infection media/BSA (DMEM supplemented with GlutaMAX, 30mM MgCl₂ and 0.2% BSA, Invitrogen). Screening plates were incubated for 7h at 37°C.

The cells were fixed by adding 33µl of 16% PFA directly to the culture medium. Fixation was either for 30min at RT or long term storage at 4°C. Cells were washed twice with PBS/25mM NH₄Cl, permeabilized with 50µl 0.2% Triton X-100 (Sigma-Aldrich) followed by 3 PBS washes and blocking with PBS containing 1% BSA (Fraction V, Sigma-Aldrich). Fixed and permeabilized cells were incubated at RT for 1h with diluted mabR16-7 antibody (0.45µg/ml) in PBS/1% BSA. Cells were washed 3 times with PBS and incubated with 25µl
secondary staining solution (PBS/1% BSA) containing Alexa Fluor 488 secondary antibody (1µg/ml, Invitrogen), DAPI (1µg/ml, Sigma-Aldrich), and DY-647-phalloidin (0.2U/ml, Dyomics). Cells were washed twice with PBS after 2h of incubation in secondary staining solution and stored in 50µl PBS/NaN₃.

**Salmonella typhimurium Protocol**

All liquid handing stages of infection, fixation, and immunofluorescence staining were performed on a liquid handling robot (BioTek; EL406). For infection the *S.typhimurium* strain S.TmSopE.pM975 was used. This strain is a single effector strain, only expressing SopE out of the main four SPI-1 encoded effectors (SipA, SopB, SopE2 and SopE). Additionally this strain harbors a plasmid (pM975) that expresses GFP under the control of a SPI2 (ssaG)-dependent promoter. The bacterial solution was prepared by cultivating a 12h culture in 0.3M LB medium containing 50µg/ml streptomycin and 50µg/ml ampicillin. Afterwards a 4h subculture (1:20 diluted from the 12h culture) was cultivated in 0.3M LB medium containing 50µg/ml streptomycin, which reached an $OD_{600nm} \approx 1.0$ after the respective 4h of incubation time. To perform the infection, 16µl of diluted *S. typhimurium* (MOI= 80) were added to the HeLa cells. After 20min of incubation at 37°C and 5%CO₂, the *S. typhimurium*-containing media was replaced by 60µl DMEM/10% FCS containing 50µg/µl streptomycin and 400µg/µl gentamicin to kill all remaining extracellular bacteria.

After additional 3h 40min incubation at 37°C and 5%CO₂, cells were fixed by adding 35µl 4% PFA, 4% sucrose in PBS for 20 min at RT. The fixation solution was removed by adding 60 µl PBS containing 400µg/ml gentamicin. Cells were permeabilized for 5 min with 40µl 0.1% Triton X-100 (Sigma-Aldrich). Afterwards 24µl of staining solution containing DAPI (1:1000, Sigma-Aldrich) and DY-547-phalloidin (1.2 U/ml, Dyomics) was added (prepared in blocking buffer consisting of 4% BSA and 4% Sucrose in PBS). After 1h of incubation at RT, cells were washed three times with PBS followed by the addition of 60µl PBS containing 400µg/ml gentamicin.
**Shigella flexneri Protocol**

*S. flexneri* M90T Δ*vir* pCK100 (PuhpT::dsRed) were harvested in exponential growth phase and coated with 0.005% poly-L-lysine (Sigma-Aldrich). Afterwards, bacteria were washed with PBS and re-suspended in assay medium (DMEM, 2 mM L-Glutamine, 10 mM HEPES). 20µl of bacterial suspension was added to each well with a final MOI of 15. Plates were then centrifuged for 1 min at 37°C and incubated at 37°C and 5% CO2. After 30 min of infection, 75µl were aspirated from each well and monensin (Sigma) and gentamicin (Gibco) were added to a final concentration of 66.7µg/ml. The total infection time was 3.5 h.

Cells were fixed in 4% PFA for 10 min after the infection phase. The liquid handling was performed using the Multidrop 384 (Thermo Scientific) and a plate washer (ELx50-16, BioTek) for aspiration steps. For immunofluorescent staining, cells were washed with PBS using the Power Washer 384 (Tecan). Subsequently, cells were incubated with a mouse anti-human IL-8 antibody (1:300, BD Biosciences) in staining solution (0.2% saponin in PBS) for 2 h at RT. After washing the cells with PBS, Hoechst (5µg/ml, Invitrogen), DY-495-phalloidin (1.2 U/ml, Dyomics) and Alexa Fluor 647-coupled goat anti-mouse IgG (1:400, Invitrogen) were added and incubated for 1 h at RT. The staining procedure was performed using the Biomek NXP Laboratory Automation Workstation (Beckman Coulter).

**Vaccinia virus Protocol**

All liquid handing stages of infection, fixation, and immunofluorescence staining were performed on a liquid handling robot (BioTek, EL406). For infection assays a recombinant WR VACV, WR E EGFP/L mCherry, was utilized. For infection, media was aspirated from the RNAi-transfected cell plates and replaced with 40 µl of virus solution per well (MOI= 0.125). Screening plates were incubated for 1 h at 37°C to allow for infection, after which virus-containing media was removed and replaced with 40µl DMEM/10% FCS. 8 h after infection 40µl of DMEM/10%FCS containing 20µM cytosine arabinoside
(AraC) was added to all wells to prevent virus DNA replication in secondary infected cells. The total infection time was 24h.

Cells were fixed by the addition of 20µl 18% PFA for 30min, followed by two PBS washes of 80µl. For immunofluorescence staining of EGFP, cells were incubated for 2h in 30µl primary staining solution (0.5% Triton X-100, 0.5% BSA, PBS) per well, containing anti-GFP antibody (1:1000). Cells were washed twice in 80µl PBS, followed by the addition of 30µl secondary staining solution (0.5% BSA, PBS) containing Alexa Fluor 488 secondary antibody (1:1000), Hoechst (1:10000), and DY-647-phalloidin (1:1200, Dyomics). Cells were washed twice with 80µl PBS after 1h incubation in secondary staining solution followed by the addition of 80µl H2O.
We propose a set of statistical methods to analyze genome-wide RNAi screens having only few replicates. These include the stable Lasso to account for the seed effects due to the siRNA technology. Furthermore, we developed a method for the analysis of the microRNA mimic screens which convolutes the readouts of miRNA mimic screens by quantifying the marginal readout effect caused by a single gene. Therefore, it enables a direct comparison between miRNA mimic screens and siRNA screens. Last, we present an approach based on Random Forest to analyze the cell behavior of the top scoring genes on the level of single cell data. The comprehensive siRNA dataset will be provided as a public, freely available resource for further statistical and biological analyses in the high-content and high-throughput siRNA screening field.
4.1 Background

The analysis of high-throughput experiments typically deals with the problem of high-dimensionality, i.e. the number of variables exceeds the number of observations ($p >> n$). This poses several problems for the detection of the relevant variables that would be interesting for follow-up experiments. One solution is the assumption of sparsity on the underlying structure. For most applications this is a reasonable assumption. In case of large-scale RNAi screens, around 22000 genes are knocked down, of which only a few actually play a role in the studied biological process. The number of replicates per gene is typically small due to time and costs.

An additional problem to the high-dimensionality is the low reproducibility of large-scale siRNA screening data due to off-target effects (Marine et al. 2012). This adds considerable noise to the observed phenotypic readouts. Thus, the usual analysis procedures, as the ranking of genes according to their raw phenotypic readout by mean or median, do not yield reliable results. A comparison of the results from siRNA-based studies shows almost no intersection between the top scoring genes (Sigoillot et al. 2012). A correction for seed effects is essential in order to reduce the number of false positives in the results and in the follow-up experiments.

Once a reliable ranking of the genes is achieved, a threshold needs to be chosen to distinguish between hit and non-hit genes. In practice, the threshold is often selected based on the capacity of re-screening. Therefore, we will focus in this chapter on how a reliable ranking is set up and how this can be combined with the results of the follow-up experiments in order to get a double-validated (by primary and secondary screen) list of interesting genes. Additionally, we provide a statistical method to analyze the cell behavior under infection on the level of individual cell data. As an example, we use the genome-wide RNAi screens of the InfectX consortium and their follow-up experiments.
4.2 Results

4.2.1 Genome-Wide RNAi Screening for Infection

The InfectX consortium performed several genome-wide siRNA screens for five bacterial and three viral pathogens (see Figure 4.1). The screens were conducted under highly unified wet-lab and analysis protocols in order to receive comparable results between the pathogens. RNA interference was achieved by commercially available siRNAs provided by two manufacturers: Dharmacon with a pool of 4 siRNAs (Human ON-TARGETplus SMARTpool) and Qiagen with 4 independent siRNA (Human Druggable Genome siRNA Set V4.0). The number of replicates differed depending on the pathogen and library (see Table 4.1).

Figure 4.1: Overview of the InfectX experiments and data analysis (A) The figure shows examples of microscope images for all eight pathogens after transfection with an siRNA and the infection phase. (B) A set of n features was extracted from each individual cell and was used for the classification of cells into infected and not infected by a decision tree algorithm. (C) The figure outlines the unified analysis work flow for all the pathogens. Only the parts that are marked in green were different between the pathogens.
As described in Section 2.4.2, HeLa cells were infected with a pathogen. This infection phase was designed by a pathogen-specific protocol so that the rate of infection and the cell number ended up as similar as possible between all pathogens (see Table 3.2). All the screened plates were imaged by automated microscopy and further processed through image analysis in order to yield a phenotypical well readout that is suited for cross-pathogen comparison after standard normalization (see Figure 4.1.B-C).

In addition to the primary genome-wide screen, the InfectX consortium performed a secondary screen on a selected set of 1520 genes (see Section 4.2.3 for details). For this we ordered up to 6 individual, commercially available siRNA and esiRNA sequences per gene. The manufacturer Ambion provided the siRNA sequences (Silencer and Silence Select) and the manufacturer Sigma the endoribonuclease-prepared siRNA (esiRNAs) pools. Furthermore, a high throughput screen was conducted with a comprehensive microRNA mimic library (Dharmacon miRIDIAN microRNA Library - Human Mimic). All secondary screening plates were processed by the same protocols as the primary screen. The same applies to image analysis, data preprocessing and normalization methods (see Section 4.3 for details).

### 4.2.2 Off-Target Correction

One challenge working with siRNAs is the correction for off-target effects. Readouts of genome-wide siRNA screens are composed of two effects: the effect induced by the on-target gene, that is the determined target of the siRNA, and the effect induced by the off-target genes, that are not intended to be knocked down by the siRNA (Birmingham et al. 2006). The siRNA behavior is comparable to miRNAs and primarily caused by the so called ”seed” region (nucleotide positions 2 to 8) (Jackson et al. 2006). We compared the readouts of different siRNA targeting the same gene in order to confirm the presence of off-target effect in our data. The normalized infection index of two different siRNAs showed a Pearson correlation coefficient between 0 and 0.1 (see Figure 4.2.B).
Figure 4.2: Off-Targets and Seed Effects (A) The seed regions are extracted from all the siRNAs and are arranged in a table which serves as input for the off-target correction method. (B) The histogram shows the Pearson correlation coefficients between different siRNAs targeting the same gene. The correlations are very small due to off-target effects. (C) The Venn diagram shows the overlap of the stable seeds between the infection index and the cell count for bacteria and viruses. (D) The figure gives an example of the seed effect for the *Vaccinia virus*. The two boxplots on the left side show the effect of the seed CGACCGC on the phenotypical readout cell count (based on 60675 observations without this seed and 131 with this seed). The two boxplot on the right show the effect of the seed ACCACAA on the infection index (based on 60774 observations without this seed and 32 with this seed).
A reduction in terms of off-target effects was achieved through the deconvolution method DecoRNAi developed by Zhong et al. 2014. DecoRNAi fits a $L_1$ regularized least-squares regression (Lasso) to the readout of RNAi experiments using a table of 7mer seed sequences as design matrix (see Figure 4.2.A). We refined this method in its weak points before we applied it to our data. In a first step, we replaced the arbitrary choice of the threshold for Lasso by the stability selection. This enabled the choice of a threshold based on the criterion of false positives. The relation between the shrinkage and the number of false positives was shown by Meinshausen and Bühlmann 2010. Secondly, we predicted the seed effects by an ordinary regression method to avoid the estimation bias that was introduced by the prediction of off-targets with Lasso. As a last step, we extended the method for the use of pooled siRNA libraries (see Section 4.3 for details).

The refined off-target correction method yielded a corrected phenotypical readout, as well as a set of seeds that caused with high probability seed-driven phenotypic readouts. An example of such a seed effect is given in Figure 4.2.D. We found that a total of 125 influential seeds are shared between the bacterial, respectively viral pathogens. The stable seeds differed between the phenotypical readouts. The number of shared seeds that provoked a bias of the readout cell count was higher compared to the readout infection index (see Figure 4.2.C). An explanation is that the readout cell count is a pathogen-independent readout, whereas the infection index bases on the pathogen-specific biological pathways leading to infection.

4.2.3 Analysis of RNAi Follow-up Screens

Compared to the primary RNAi screen, follow-up screens typically consider only a few hundred genes. Another difference is that the number of replicates is much higher since the time to produce another full assay is no longer a limiting factor. The InfectX consortium selected a set of 1520 genes for the follow-up screens. A gene was included into the selection if one of the following three criteria matched:
• The gene appeared on the top or the bottom of the ranking of the normalized and seed-corrected infection index readout.

• The gene was enriched in KEGG pathways related to pathogen entry and infection.

• The gene got randomly selected as control.

If a gene was selected for one pathogen, then it was screened for all the pathogens. The follow-up screens were replicated up to 6 times.

Through the increased number of replicates we can benefit from the fact that the measured response of the intended knocked down gene is on average enhanced, since the independent siRNAs replicates should differ in their off-targets, but not in their on-target (Bassik et al. 2013). In Section 3.2.3 we presented the Parallel Mixed Model (PMM) which exploits this assumption. Moreover, PMM gains statistical power for the detection of genes shared between the studied conditions (e.g. pathogens, cell lines etc.) by modeling them simultaneously. Even if the studied conditions show only a slight overlap, PMM can improve the hit detection in individual screens without loss for the detection of non-shared hits (see Section 3.2.6).

The secondary siRNA and esiRNA screens were thus analyzed with PMM. The normalized infection index readout was used as input for PMM. We added a-priori information about the performance of the RNAis into the PMM and gave double weight to measurements of esiRNAs, since esiRNA are known to be more on-target specific and to have less off-target effects (Myers et al. 2006). The PMM provided an estimation for the effect of a gene within a certain pathogen, as well as a false discovery rate for the estimated effect. A gene was said to be significant if its false discovery rate was below 0.2.

4.2.4 Deconvolution of MicroRNA Mimic Screen

MiRNA mimics have been designed to mimic the miRNA molecules. They down-regulate the expression of the targeted genes. Identically
to the siRNAs, the miRNA mimics are transfected into cells and then a phenotypical response of the cell is measured. The biologists can infer from these experiments a set of regulatory components. Usually biophysical model are used to predict if a certain gene is contained in the target set of an miRNA. Our developed deconvolution method goes one step further and quantifies how much a single gene contributed to the phenotypical readout of an miRNA. This enables a direct comparison of miRNA and siRNA screens and the use of miRNA mimic screens as confirmation for primary RNAi screens.

Our deconvolution technique is based on the biophysical model MIRZA-G (Gumienny and Zavolan 2015). MIRZA-G predicts all the possible targets for an miRNA through a logistic regression model. These predicted targets serve us for the estimation of the marginal gene effect. We arranged the predicted target probabilities in a matrix $x$ where each row corresponds to an miRNA and one column to a gene. The phenotypical response $y_i$ of one miRNA $i$ is modeled as weighted sum of all genes:

$$y_i = \sum_j x_{ij} \beta_j + \epsilon_i.$$  

The coefficients $\beta_j$ are estimated with stable Lasso (see Section 4.3.5 for details). The $\hat{\beta}_j$ represent the effect of a gene on the phenotypic readout. To benefit from the parallel screening of $p$ pathogens, we fitted two Lasso regressions: one with a one-dimensional response and the other with a $p$-dimensional response matrix. The two resulting estimated coefficients were averaged and ranked. All genes corresponding to non-zero coefficients belong to the set of genes having an influence on the phenotypic response.

### 4.2.5 Detection of Relevant Genes

Applying the previously presented methods to the infection index readout of the InfectX RNAi screens, we obtain three results for the prioritization of genes according to infection. The primary screen
Figure 4.3: Summary of Detected Hits. The heat map lists all the genes that were confirmed by a secondary screen at least for one pathogen. The colors correspond to the averaged infection score. Hits are marked with a yellow, green or black symbol depending on how they were confirmed. NA indicates that there was no measurement of the corresponding gene in the analysis. The genes are sorted by the amount of sharing between the pathogens and the strength of the infection rate.
yields a gene ranking based on the mean of the normalized, seed-
corrected infection index calculated over all siRNA targeting a cer-
tain gene. From the secondary siRNA and esiRNA screens we obtain
through PMM a total of 188 genes that have false discovery rate be-
low 0.2. As third result, we obtain in total 450 genes with non-zero
coefficients from the miRNA screen. In a next step we present an
approach about these results can be combined into a reliable list of
relevant genes.

We calculate a threshold for the primary ranking using the hyperge-
ometric distribution. The hypergeometric distribution describes the
probability of having k successes in n draws from a given population.
The primary screen defines the population with its N genes and K
hits and the secondary screens yields m interesting genes from the
original N genes, of which x genes are contained in K. The missing
parameter K of the hypergeometric distribution was defined by the
condition that the probability of obtaining x hits was maximized and
not allowed to be in the outer 1% of the hypergeometric distribution
for all pathogens. For the InfectX RNAi screens we found that the
threshold should be $K = 1284$ (taking this threshold from both sides
of the ranking).

We define as relevant all the genes which are above the threshold and
which are confirmed by at least one secondary screening result. In
total, we identified 216 genes as influential on the infection index (see
Figure 4.3). About one third of these genes are shared between the
pathogens. The amount of shared genes is higher among the viral
pathogens than the bacterial pathogens. This can be motivated by
the fact that there are more distinct bacterial entries than viral entry
mechanism. Adenovirus and Rhinovirus enter both by clathrin- and
dynamin-dependent pathways.

4.2.6 Study of Phenotypic Patterns

The image analysis provided about 300 features per gene knockdown
for the several hundred individual cells in a well. From each indi-
individual cell the geometry, textures and intensities of the nuclear and cellular objects, as well as neighboring features were extracted using the actin and DNA stain values (see Section 2.4.2). So far we used only a small set of these features to construct the infection score. Comparing these additional cell features between infected and not infected cells should reveal more details on the infection phenotype under the knockdown of a certain gene. From the set of detected hit genes (see Figure 4.3) we took the six top shared genes (\textit{ACTR3, NPAT, PIGH, TLN1, ITGB1} and \textit{ZEB1}) in order to enable cross-comparison of the results between the pathogens.

We used the Random Forest (Breiman 2001) on the individual cell features to predict the infected cells for each well. Random Forest was chosen because it can handle mixed-type data, it corrects for overfitting and it is known to perform very well under complex interactions and non-linear data structures. We obtained from the Random Forest a well-based ranking of the features by their importance score. These single rankings were averaged to a ranking per gene. We determined a threshold for the ranking by cross-validation. The number of correctly predicted cells served as cross-validation criterion (see Section 4.3.6 for details). A corresponding logistic regression model confirmed the selected features as significant.

Figure 4.4 illustrates the results of the single cell analysis. The list contains several genes of previously reported pathways. For instance, \textit{ITGB1} is a member of the integrin signaling pathway and triggers major rearrangement of the actin for the formation of invasomes in the case of \textit{B. henselae} (Truttmann et al. 2011) or \textit{ACTR3} is known as regulator of the actin polymerization for \textit{S. typhimurium} (Unsworth et al. 2004). The single cell features confirm this findings. For both genes the actin channel of the texture and intensity are very predictive for infection. \textit{B. abortus} gives another evidence that this analysis provides interesting features. Nearly all the features are important. This corresponds perfectly with the activity of the pathogen. \textit{B. abortus} causes disturbance across the whole cell body with its wide spread micro-colonies.
Figure 4.4: **Single Cell Analysis.** The heat map shows the predictive power of features group on infection for the top shared and influential genes. The colors correspond to the averaged importance score per features group. Higher values mean more important. White means that none of the features in the corresponding feature group was relevant.
4.3 Methods

In this methods section we give details how the InfectX data were pre-processed and explain the statistical methods that were presented in the previous section.

4.3.1 Data Preprocessing

For the prediction of siRNA targets, we mapped the siRNA sequences against the genomic transcript sequences from RefSeq (release hg19) and ENSEMBL (release GRCh37.67) using BLAST version 2.2.27. The obtained transcript IDs were then translated to Entrez gene IDs whenever both obtained transcript IDs agreed. Furthermore, we removed all the wells and all the plates that were flagged as experimentally biased during the experiments. We also removed all the wells that had less cells than the number of seeded cells in order to avoid unwanted effects of cell toxicity from the siRNA (see Section 4.B for more details).

4.3.2 Data Normalization

We applied the B-Scoring algorithm (Brideau et al. 2003) to the raw infection index in order to remove systematic within-plate effects. The B-Scoring algorithm estimates the within-plate bias by fitting a two-way median polish to the rows and the columns of each 384-well plate. The corrected data corresponds to the residuals, that is the raw infection index minus the overall infection index on the plate, the row effect and the column effect. Furthermore, we used Z-Scoring in order to eliminate experimentally introduced cross-plate biases. Therefore, the B-scored infection indices of each plate were centered by the mean of the plate and scaled by the standard deviation of the plate. At last, we averaged the infection index from all wells that dealt with siRNAs having identical seed sequence and with the same pathogen.
4.3.3 Seed Effects

As seed correction method we applied a refined version of the deconvolution analysis of RNAi screening data (DecoRNai) (Zhong et al. 2014) to our data. First, we extracted from all siRNA oligonucleotides the seed region (nucleotide positions 2 to 8), as well as a shifted version (nucleotides positions 1 to 7 and nucleotides positions 3 to 9). From this collection of seeds we removed the seeds which appeared only once and we constructed a design matrix $X$ with the following entries

$$x_{ij} = \begin{cases} 
1, & \text{if siRNA } i \text{ contains seed } j \\
0, & \text{otherwise.} 
\end{cases}$$

For the pool of four siRNA, we set $x_{ij} = 1$ if one of the siRNAs in the pool $i$ contained seed $j$. Next, we fitted a $L_1$ regularized least-squares regression (Lasso) using $X$ as design matrix. The response vector was given by the normalized infection index per well where the $i$th entry denotes the infection index corresponding to a knockdown with siRNA $i$ or the siRNA pool $i$. The coefficients of the Lasso regression describe the off-target effect caused by a certain seed. Relevant off-target effects were selected by the stability criterion (see section 4.3.4). We bounded the number of expected false positives to 100 and set the stability parameter to 0.75. The seed-corrected infection index corresponded to the residuals. Since the coefficients estimated by Lasso are biased towards zero due to the shrinkage, we re-calculated the residuals by the ordinary least squares regression:

$$Y_{\text{cor}} = Y - X_{\text{sel}} \hat{\beta}_R$$

with the estimated coefficients $\hat{\beta}_R$ from the linear regression model using only the selected seed variables $X_{sel}$.

4.3.4 Lasso with Stability Selection

The $L_1$ regularized least-squares regression (Lasso) differs from the ordinary least squares regression model by an additional penalty
term:
\[ \hat{\beta} = \arg \min_{\beta} \left( \| Y - X\beta \|_2^2 + \lambda \| \beta \|_1 \right). \]

\( Y \) denotes the response vector, \( X \) is the design matrix, \( \beta \) is the vector of coefficients to be estimated and \( \lambda \) is the penalty parameter (Tibshirani 1996). For stability selection we repeatedly subsampled 70% of the full data and refitted the Lasso for each sampled data set. We chose all the variables as relevant which exceeded a given stability threshold \( \pi_{th} \) by their selection probability computed over all subsampling rounds.

The choice of regularization during the Lasso fit can be linked to the amount of false positives (Meinshausen and Bühlmann 2010). The expected number \( V \) of false positives among \( q(\lambda) \) selected variables from a total of \( p \) variables is bounded by
\[
E[V] \leq \frac{q(\lambda)^2}{(2\pi_{th} - 1)p}.
\]

For all the fitted Lasso models with stability selection we used the following setting (unless otherwise stated): 500 subsampling rounds, stability threshold as 0.75 and the number of false positives as 100.

### 4.3.5 Analysis of MiRNA Screens

First, we mapped the miRNA phenotypic readouts to genes. Therefore, we used a biophysical model to predict the possible targets for miRNA, called MIRZA-G (Gumienny and Zavolan 2015). MIRZA-G uses a generalized linear model with a logit link function. It is trained on data from 26 experiments measuring the gene expression level while transfection with miRNA. As covariates MIRZA-G uses the target quality score (Khorshid et al. 2013), which measures the miRNA target interaction energy between an mRNA and miRNA predicted by MIRZA, the position of the target site in 3' UTRs, the nucleotide content and the accessibility of the target site. We saved the predicted probabilities of MIRZA-G in a matrix \( x \) where each row
corresponds to an miRNA and one column to a gene and modeled the phenotypical response of one miRNA as weighted sum of all genes:

\[ y^p_i = \sum_j x_{ij}\beta^p_j + \epsilon_i \]

We selected the relevant gene effects \( \beta^p_j \) for each pathogen \( p \) by eight separate stability lasso regressions (see section 4.3.4). The strength of the selected genes \( \beta^p_j \) was estimated by linear regression containing as covariates only the selected genes. Additionally, we fitted a multivariate lasso regression using a \( p \) column length matrix with phenotypes of all pathogens as response. Thus, for each pathogens we obtained two estimated effects, one from the single Lasso and one from the multivariate. We combined the phenotypical effects by taking the mean.

### 4.3.6 Single Cell Analysis

The single cell features were first normalized by subtracting the mean value of the well. Then we fitted Random Forests (Breiman 2001) separately to the features of all cells in one well. The Random Forest provided an important measure for each feature. The importance measure is based on the out-of-bag error which is estimated by leaving one-third of the cases out of the bootstrapped samples. We averaged the importance scores over all the wells containing cells which were knocked down by a siRNA targeting the same gene. This allowed us to receive an overall ranking of the features for a certain gene.

The threshold for the ranking was defined by the maximum of the mean number of correctly predicted cells. This maximum was calculated over \( k \) cross-validation using the top \( k \) ranked features as input for the Random Forest. As training and test data served all the combinations of paired wells. In each cross-validation round one well was chosen as training set and the other well as test set. If the maximum rate of correctly predicted cells was below 0.5, the model was said to be not significant. For fitting we used the R package \texttt{randomForest} (Liaw and Wiener 2002).
4.4 Discussion and Conclusions

One challenge in high-content screening is clearly to distinguish important genes from the noise caused by the massive amount of data. If a gene appears in two independent screens as top ranked, then the probability that we found an important gene is much higher. With the approach of deconvoluting miRNA mimic screens, we enriched the large-scale screens with additional independent observations. One miRNA knockdown covers thousand of genes and therewith saves a lot of time and cost. Moreover, by the hypergeometric bound we gave a flexible way to find significant genes without using an a-priori threshold on the ranking of the primary screen. With our approach we found 216, among them also novel, genes that are influential on infection.

We proposed a refined method for seed-effect correction, which has the advantage of false positive control. However, it is challenging to show that the off-target correction is successful, since there is still an insufficient knowledge on the off-target behavior of individual siRNAs. Published papers about off-target correction motivate their predictions through the correlation of independent siRNA experiments by claiming that the average signal corresponds more to the true on-target signal, or they give an increased biological enrichment of the top ranked genes (Zhong et al. 2014, Franceschini et al. 2014, Schmich et al. 2015). Recently, a new technology named CRISPR/Cas9 (Cong et al. 2013) was developed. This technology enables a clean knock-off of a certain gene. However, it is quite expensive and not yet available for high-throughput screens.

We analyzed the cell behavior of the top hits in the genome-wide screen on the level of individual cell data by applying Random Forests. We could show that there is additional information in the individual cell features. If gene pathways are of interest, then the analysis could be applied to the individual cell data of genes within the interested pathway. The resulting features could be clustered and the groups could give more insight into the functions within the pathway.
Appendix Chapter 4

4.A Number of Replicates

The following table reports the number of replicates for each pathogens and RNAi screen.

<table>
<thead>
<tr>
<th>Pathogen</th>
<th>Primary Screen</th>
<th>Secondary Screen</th>
<th>miRNA Screen</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dharmacon 1 pooled siRNA</td>
<td>Qiagen 4 unpooled siRNA</td>
<td>Ambion 6 unpooled siRNA</td>
</tr>
<tr>
<td>Adenovirus</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>B. abortus</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>B. henselae</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>L. monocytogenes</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Rhinovirus</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>S. flexneri</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>S. typhimurium SipA</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>S. typhimurium SopE</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Vaccinia virus</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 4.1: Replicates performed for different pathogens and libraries.

^1 The replicates differ in the concentration of added miRNA.

4.B Removing Cell Killers

The following table shows the number of seeded cells per pathogen:

<table>
<thead>
<tr>
<th>Pathogen</th>
<th># Seeded cells per well</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenovirus</td>
<td>700</td>
</tr>
<tr>
<td>B. henselae</td>
<td>300</td>
</tr>
<tr>
<td>B. abortus</td>
<td>500</td>
</tr>
<tr>
<td>L. monocytogenes</td>
<td>600</td>
</tr>
<tr>
<td>Rhinovirus</td>
<td>1000</td>
</tr>
<tr>
<td>S. Typhimurium SipA/SopE</td>
<td>550</td>
</tr>
<tr>
<td>S. flexneri</td>
<td>600</td>
</tr>
<tr>
<td>Vaccinia virus</td>
<td>600</td>
</tr>
</tbody>
</table>

A high degree of cell death is induced in wells containing cells in which a gene that is essential for cell survival is knocked down by a siRNA. For the analysis of infection, cell killing siRNAs are not of interest.
Figure 4.5: **Cell Killers.** The left plot shows distribution of cell counts relative to the seeded number for all wells of the primary screen. The right plot shows the same graphics for all wells of the secondary siRNA and esiRNA screen.

Therefore, we excluded all the wells from the analysis which had a cell count below its seeded number. Figure 4.5 shows the distribution of cell counts relative to the seeded number for all the wells of the primary screen and all the wells of the secondary siRNA and esiRNA screen.
In siRNA screens the observed phenotypic readouts are biased due to the fact that siRNAs not always bind to their intended target. A direct reconstruction of the gene network from the experimental readout data is therefore not possible. We propose an algorithm that exploits the off-target effects as multiple knockdowns and reconstructs from them an estimate for the inverse covariance matrix of the pure gene effects. Our estimation procedure is based on the EM-algorithm and graphical Lasso. It assumes that the off-target relationships are known. The estimated inverse covariance matrix enables the reconstruction of a gene network, for example by using it as input to the PC-algorithm.
5.1 Background

Many biological processes involve a set of proteins, as for example the progression through the cell cycle or cellular uptake of particles. The interaction between proteins plays an essential role for human life in order to smoothly perform complex cellular tasks. Through RNAi experiments, only information on a single gene is typically deduced from the readout data. However, in order to comprehensively understand a cellular process, the full knowledge on the interaction between the genes is necessary.

Knowing the inverse covariance matrix between genes is a first step towards the estimation of gene networks, since it reflects, under certain assumptions, the conditional independence structure of graphs. There are several approaches that successfully extract information about an underlying graph from the inverse covariance matrices. One example is the parametrization as graphical Models (Lauritzen 1996), or the PC algorithm which builds a directed acyclic graph based on conditional independence tests (Sprites et al. 2000). Therefore, we aim here to estimate the inverse covariance matrix from RNAi screening data.

The direct estimation of the inverse covariance matrix is challenging due to off-target effects. They arise when an siRNA partially shares its base sequence with other genes. However there are algorithms, which can predict all the possible targets of an siRNA by comparing the base sequences, the gene expressions and including other possible binding information (e.g. Target Scan (Lewis et al. 2005) or the biophysical model MIRZA-G (Gumienny and Zavolan 2015)). Thus, we propose here an algorithm to estimate the partial correlation between the genes under the assumption that the off-target relationships are known. We confirm the performance of our proposed algorithm through a simulation study.
5.2 Underlying Model

Let the random variable $Y_j$ ($j = 1, \ldots, p$) describe the phenotypic readout that is obtained through an siRNA knockdown experiment with gene $j$ ($j = 1, \ldots, p$) as target. Due to off-target effects the observed phenotypic readout $Y_j$ is a combination of the phenotypic response from the intended target gene $j$ and from other genes that are partially targeted by the siRNA. Through algorithms, as for example the biophysical model MIRZA-G (Gumienny and Zavolan 2015), we can predict the probabilities that a certain gene $i$ is affected by the knockdown of siRNA $j$. We store these contribution $w_{ij}$ of a single gene $i$ to the observed phenotypic readout $Y_j$ in a matrix $W$. As last, we define the random variables $G_k$ ($k = 1, \ldots, p$) which describe the phenotypic readout of a clean knockdown experiment, i.e. the siRNA binds only to the mRNA of gene $k$. Figure 5.1 illustrates the connection between the defined variables $G$ and $Y$.

![Diagram](image)

Figure 5.1: The vertices on the left side represent the genes and the vertices on the right side the observed phenotypic responses. Due to off-target effects there is no one-to-one relation. Each observed phenotypic response originates from the simultaneous (but not equally strong) knockdown of several genes. The weight on the edges refers to the different contributions. In addition, the edges on the left illustrate the interaction between the genes.
If an siRNA would only target the gene that it was designed for, then the relation between $Y$ and $G$ would look as follows

$$Y_j = G_j, \quad \forall j = 1, \ldots, p$$

$$w_{ij} = \begin{cases} 
1 & \text{if } i = j \\
0 & \text{if } i \neq j.
\end{cases}$$

This relation changes to a weighted sum if we account for the off-targets

$$Y_j = \sum_{k=1}^{p} w_{kj} G_k,$$  \hspace{1cm} (5.1)

where $w_{kj}$ corresponds to the contribution of gene $k$ on the phenotypic response $Y_j$. The weight $w_{kj}$ is equal 0 if gene $k$ is not in the set of off-target genes of siRNA $j$.

Without loss of generality we will assume that the weights in each row sum up to 1

$$\sum_{k=1}^{p} w_{kj} = 1, \quad \forall j.$$  \hspace{1cm}

Otherwise, the matrix $W$ can be scaled so that this assumption holds. In addition, we assume that the phenotypic behavior of the genes is normally distributed

$$(G_1, \ldots, G_p) \sim \mathcal{N}(\mu, \Sigma_G).$$  \hspace{1cm} (5.2)

From this we can derive that the observed phenotypic response is normally distributed, too

$$(Y_j | G_1, \ldots, G_p) \sim \mathcal{N}(\sum_{k=1}^{p} w_{kj} G_k, \sigma^2).$$

Furthermore, $(Y_k | G_1, \ldots, G_p)$ and $(Y_j | G_1, \ldots, G_p)$ are conditionally independent for $k \neq j$. 
5.3 Recovery of the Gene Effects and their Concentration Matrix

Our aim is the estimation of the covariance matrix $\Sigma_G$, respectively the concentration matrix $K := (\Sigma_G)^{-1}$. For this estimation it is necessary to know the unobserved gene effects $G$. Therefore, we derive in a first step an estimator for the gene effects $G$ and in a second step an estimator for the concentration matrix $K$. Then, we combine the two estimators iteratively in order to optimize the estimations.

5.3.1 Estimation of the Gene Effects

Let $y_i = (y_{1i}, \ldots, y_{pi})^T$ be sample data ($i = 1, \ldots, n$) on the phenotypic readout from an siRNA screen with $p$ single knockdowns, each having a different gene as targets. A good estimator for the gene effects $G$ is clearly

$$\hat{G} = (\hat{G}_1, \ldots, \hat{G}_p) = E[G \mid y, W]$$

where $W = (w_{ij})_{i,j=1,\ldots,p}$ refers to the off-target matrix and $y = (y_1, \ldots, y_p)$ denotes the sample matrix. From the two assumptions (5.1) and (5.2) we can derive a statement on the conditional distribution of $(G \mid Y)$. Given the phenotypic readouts, the pure gene effects are normally distributed with expectation

$$E[G \mid y] = (K + \frac{1}{\sigma^2}WW^T)^{-1}(K\mu + \frac{1}{\sigma^2}Wy).$$

A proof is given in the Appendix (see Lemma 5.A.1). Hence an estimate for the unobserved $G$ is given by

$$(\hat{G}_1, \ldots, \hat{G}_p) = (K + \frac{1}{\sigma^2}WW^T)^{-1}(K\mu + \frac{1}{\sigma^2}Wy).$$

Note that its calculation requires besides the data $y$ the knowledge of the parameters $\sigma$, $K$ and $\mu$. 
5.3.2 Estimation of the Concentration Matrix

In this section we derive an estimator for the parameters $\sigma$, $K$ and $\mu$. We assume for this section that we are able to observe sample data $g_i = (g_{1i}, \ldots, g_{pi})^T$ for the pure gene effect. The complete log-likelihood is given by

$$l(\mu, \Sigma_G, \sigma; g, y) = -np \log(2\pi) - \frac{n}{2} \log(|\Sigma_G|)$$

$$- \frac{1}{2} \sum_{i=1}^{n} ((g_i - \mu)^T \Sigma^{-1}_G (g_i - \mu))$$

$$- \frac{n}{2} \log(\sigma^2 p) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - W^T g_i)^T (y_i - W^T g_i).$$

For the derivation of the joint distribution of $Y$ and $G$ we refer to Section 5.A. Since there is biological evidence that genes networks are sparse (Thieffry et al. 1998), we add a penalty to the likelihood estimation of $K$. The penalized log-likelihood is then given by

$$l_\lambda(\mu, K, \sigma; g, y) = -np \log(2\pi) + \frac{n}{2} \log(|K|) - np \log(\sigma)$$

$$- \frac{1}{2} \text{tr} \left( K \left[ \sum_{i=1}^{n} (g_i - \mu)(g_i - \mu)^T \right] \right)$$

$$- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - W^T g_i)^T (y_i - W^T g_i) - \lambda \|K\|_1$$

where $\|K\|_1 = \sum_{k,l=1}^{p} |K_{kl}|$ and $\lambda$ is the regularization parameter, which controls the amount of penalization.

The maximum likelihood estimators of the complete likelihood can
be easily computed for $\mu$ and $\sigma$:

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} g_{ki}, \quad \text{for } k = 1, \ldots, p$$

$$\hat{\sigma}^2 = \frac{1}{np} \sum_{i=1}^{n} (y_i - W^T g_i)^T (y_i - W^T g_i).$$

In the case of $K$ the optimization problem can be reformulated to

$$\hat{K} = \arg \max_{K > 0} \left( \log |K| - \text{tr}(KS) - \lambda \|K\|_1 \right) \quad (5.3)$$

with

$$S = \frac{1}{n} \sum_{i=1}^{n} (g_i - \hat{\mu})(g_i - \hat{\mu})^T.$$

Friedman et al. 2007 proposed an optimization algorithm called Graphical Lasso (Glasso) which efficiently solve the problem (5.3). We briefly give an idea of their coordinate-wise algorithm. Details can be found in Banerjee et al. 2008 and Friedman et al. 2007.

Let $V$ be an estimate of $\Sigma$. We partition $V$ and $S$ in blocks as follows:

$$V = \begin{pmatrix} V_{11} & v_{12} \\ v_{12}^T & v_{22} \end{pmatrix}, \quad S = \begin{pmatrix} S_{11} & s_{12} \\ s_{12}^T & s_{22} \end{pmatrix}.$$

where $v_{12}$ respectively $s_{12}$ represent the last column without the entry $v_{22}$ resp. $s_{22}$. Then we can formulate the optimization problem (5.3) as lasso problem:

**Theorem 5.3.1.** The solution that is obtained from the convex optimization problem

$$\hat{K} = \arg \max_{K > 0} \left( \log |K| - \text{tr}(KS) - \rho \|K\|_1 \right)$$

for the last column or row is equivalent to

$$\hat{\beta} = \arg \min_{\beta} \left( \frac{1}{2} \| V_{11}^{1/2} \beta - V_{11}^{-1/2} s_{12} \|^2 + \rho \| \beta \|_1 \right)$$

$$v_{12} = \frac{1}{2} V_{11} \hat{\beta} \quad \text{and} \quad v_{22} = s_{22} + \rho$$
The proof can be found in Section 5.B. Theorem 5.3.1 provides a column-wise solution for the optimization problem. Through permutation of rows and columns, such that the target column is always the last one, we can iteratively update all \( k = 1, \ldots, p \) columns. Although the algorithm solves the problem for \( \Sigma \), the corresponding estimate for \( K \) can be easily obtained. The algorithm is implemented in the R-package Glasso (Friedman et al. 2014).

5.3.3 Algorithm

We iteratively combine the derived estimators of the last two sections in one algorithm. The EM algorithm gives an ideal framework for this (see Section 5.C). It also provides a proof of convergence (Dempster et al. 1977). The complete algorithm looks as follows:

**Input:** 1. Data matrix \( y \) with phenotypic readouts
2. Off-Target Matrix \( W \)
3. Initial estimates for \( \sigma \), \( K \) and \( \mu \)

**Algorithm**

**Recovery of Gene Effects and their Concentration Matrix**

1. Set \( m = 0 \).
2. Save the initial estimates as \( \sigma^{(0)} \), \( K^{(0)} \) and \( \mu^{(0)} \).
3. Iterate 4 - 12 until convergence.
4. **Iteration Step** \((m + 1)\):
   5. **E-Step**
      \[
      \hat{g}(r+1)_{p \times n} = (K^{(r)} + \frac{1}{(\sigma^{(r)})^2} WW^T)^{-1}(K^{(r)} \mu^{(r)} + \frac{1}{(\sigma^{(r)})^2} Wy)
      \]
   6. **M-Step**
      \[
      \hat{\mu}^{(r+1)}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{g}_{ki}^{(r+1)}
      \]
      \[
      \hat{K}^{(r+1)} = \text{glasso} (S_n(\hat{g}^{(r+1)}), \lambda)
      \]
      \[
      \hat{\sigma}^{(r+1)} = \frac{1}{np} \sum_{i=1}^{n} (y_i - W^T \hat{g}_i^{(r+1)})^T (y_i - W^T \hat{g}_i^{(r+1)})
      \]
5. Set \( m = m + 1 \).
The tuning parameter \( \lambda \) in the Glasso Step is selected based on the BIC criterion

\[
BIC = -2l(\hat{\mu}, \hat{K}, \hat{g}) + \log(n)k
\]

over a grid of candidate values for \( \lambda \). The two quantities \((\hat{\mu}, \hat{K})\) denote the Glasso estimates under a certain \( \lambda \). The value \( k \) corresponds to the number of non-zeros in \( \hat{K} \) and \( l(\hat{\mu}, \hat{K}, \hat{g}) \) denotes the observed log-likelihood.

5.4 Simulations

In this section we present the results of simulations to show the efficiency of our algorithm. We randomly generated possible readout data and estimate the pure gene effects, as well as their concentration matrix. Several simulation setting were considered, all were repeated \( N = 100 \) times and the results were averaged. All simulations were carried out with the statistical computing language and environment R (R Core Team 2015).

Generating Gene Effects

We sampled data from a \( p \) dimensional multivariate normal distribution with \( p \in \{5, 10, 30, 50, 100\} \), mean \( \mu \) and covariance matrix \( \Sigma = K^{-1} \). The number of replicates was chosen to be once \( n = 10 \) and once \( n = 100 \). The mean was set to zero in all the simulations. For the concentration matrix, we considered 3 models with different amount of sparsity:

**Model 1:** AR(1), \( K_{ij} = 1_{\{|i-j|=0\}} + 0.8 1_{\{|i-j|=1\}} \).

**Model 2:** AR(4), \( K_{ij} = 1_{\{|i-j|=0\}} + 0.8 1_{\{|i-j|=1\}} + 0.7 1_{\{|i-j|=2\}} + 0.6 1_{\{|i-j|=3\}} + 0.5 1_{\{|i-j|=4\}} \).
**Model 3:** We generated a matrix with randomly filled off-diagonals. Therefore, we randomly selected through a Bernoulli($\pi$) distribution whether an off-diagonal entry should be equal to zero or not. We set $\pi = 0.25$. The value of the off-diagonal entry was determined by a uniform distribution on the half-closed interval $(0, 1]$.

**Model 4:** The gene effects were generated by the same set-up as in model 3 with $\pi = 0.5$.

**Model 5:** The gene effects were generated by the same set-up as in model 3 with $\pi = 0.1$.

If a generated concentration matrix was not invertible, we added a noise vector to the diagonal entries. An example for each model is given in the first column of Figure 5.2.

**Generating Off-Target Relations**

The matrix containing the off-target relations was generated as a random matrix. We randomly selected through a Bernoulli($\theta$) distribution whether a matrix entry should be equal to zero or not. The value of the entries was determined by a uniform distribution on the half-closed interval $(0, 1]$. $\theta$ was set to different values in our simulations, in order to mimic siRNAs with strong off-target and siRNAs with less off-target activity.

**Generating Readout Data**

The readout data $Y$ were calculated by multiplying the generated gene effects with the off-target relation matrix and adding normally distributed noise with mean 0 and standard deviation $\sigma$. The value of $\sigma$ was set to different values during the simulations.
Figure 5.2: **Illustration of Simulated Models and their Estimation.** The first column shows heatmaps of the entries in the simulated concentration matrices. White color stands for a zero; grey to black refer to the non-zero entries. The second column shows the entries of the estimated concentration matrices. The last column shows a scatter plot of the true and estimated gene effects. The underlying model was generated with $n = 100$, $p = 30$, $\theta = 0.2$ and $\sigma = 0.1$. 
Measurements of Goodness

We computed three measures in order to access the performance of the different simulated models. The first two measures were the true positive rate
\[ TPR = \frac{\sum_{ij} (\hat{K}_{ij} \neq 0 \& K_{ij} \neq 0)}{\sum_{ij} K_{ij} \neq 0} \]
and the false positive rate
\[ FPR = \frac{\sum_{ij} (\hat{K}_{ij} = 0 \& K_{ij} = 0)}{\sum_{ij} K_{ij} = 0} \].

As third measure we calculated the average correlation between the true gene effects \( g \) and the estimated gene effects \( \hat{g} \).

Simulation Results

Figure 5.2 illustrates the simulated concentration matrices and their estimation. The underlying parameter were chosen as \( n = 100, p = 30, \theta = 0.2 \) and \( \sigma = 0.1 \). From the heatmaps we can see that the BIC criterion has the tendency to select a tuning parameter \( \lambda \) which yields high number of false positives and a low number of negatives. The EM-Algorithm converged typically in less than 10 steps.

We simulated four different scenarios for each concentration matrix and number of variables:

S1: \( n = 100, \theta = 0.2 \) and \( \sigma = 0.1 \).

S2: Strong Off-Target activity, \( n = 100, \theta = 0.5 \) and \( \sigma = 0.1 \).

S3: Small number of replicates, \( n = 10, \theta = 0.2 \) and \( \sigma = 0.1 \).

S4: Larger noise in \( Y \), \( n = 50, \theta = 0.2 \) and \( \sigma = 1 \).

The results are reported in Figure 5.3.
Figure 5.3: **Simulation results.** Each row shows the results for the defined models under different scenario, e.g. strong off-target activity or very small number of replicates. The first column refers to the performance according to true positive rate, the second column according to true negative rate and the last column shows the correlation between the true gene effects $G$ and the estimated gene effects $\hat{G}$. 
5.5 Discussion and Conclusions

The results on simulated data confirm that our algorithm can recover the pure gene effects and their inverse covariance matrix from the observed readout data of RNAi screens. It also indicates that the goodness of the results does not strongly depend on the off-target activity. TPR, TNR and the correlation show only minor changes between the 20% filled off-target matrix (S1) and the 50% filled off-target matrix (S2). Furthermore, we obtain promising results under high-dimensional settings, e.g. by reducing the number of replicates to $n = 10$ (S3). However, the recovery of the pure gene effects $G$ is weak if the noise in $Y$ is too large (S4).

Our algorithm has one minor drawback. For large $p$ the computation is slow, since the calculation of the Glasso estimator has a runtime which is polynomial in $p$. Since only a small subset of genes is usually involved in a biological process, one could restrict to this subset and feed the algorithm with a suitable sub-matrix $W$ and sub-sample $y$. Another possibility is the substitution of the Glasso estimator with a computationally more efficient estimator, for example with the algorithm QUIC proposed by Hsieh et al. 2012, which assumes a block-structured $K$. 

Appendix Chapter 5

5.A Distributional Statements

Joint Distribution of G and Y

Let $G = (G_1, \ldots, G_p) \sim \mathcal{N}(\mu, \Sigma_G)$ and $Y$ be defined through $Y_j = \sum_{k=1}^{p} w_{kj} G_k$ for $j = 1, \ldots, p$ and $\text{Var}(Y) = \sigma^2 I$.

Then, the expectations for $Y_i$ ($i = 1, \ldots, p$) is given by

$$E[Y_i] = E[E[Y_i \mid G_1, \ldots, G_p]] = E\left[\sum_{k=1}^{p} w_{ki} G_k\right] = \sum_{k=1}^{p} w_{ki} \mu_k.$$ 

and the covariance between $Y_i$ and $G_k$ ($i, k = 1, \ldots, p$) by

$$\text{Cov}(Y_i, G_k) = E[\text{Cov}(Y_i, G_k \mid G_k)] + \text{Cov}(E[Y_i \mid G_k], E[G_k \mid G_k])
= 0 + \text{Cov}\left( \sum_{l \neq k, l=1}^{p} w_{li} \mu_l + w_{ki} G_k, G_k \right)
= \text{Cov}\left( \sum_{l \neq k, l=1}^{p} w_{li} \mu_l, G_k \right) + \text{Cov}(w_{ki} G_k, G_k)
= 0 + w_{ki} \text{Cov}(G_k, G_k) = w_{ki} (\Sigma_G)_{kk}.$$ 

Therefore, the joint distribution $f_T$ is equal to:

$$f(t) \sim \mathcal{N}_{2p} \left( \begin{pmatrix} \sum_{k=1}^{p} w_{k1} \mu_k \\ \vdots \\ \sum_{k=1}^{p} w_{kp} \mu_k \\ \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}, \begin{pmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \\ w_{11}(\Sigma_G)_{11} & \cdots & w_{1p}(\Sigma_G)_{11} \\ \vdots & \ddots & \vdots \\ w_{p1}(\Sigma_G)_{p1} & \cdots & w_{pp}(\Sigma_G)_{pp} \end{pmatrix} \right) \begin{pmatrix} \sigma^2 \\ \vdots \\ \sigma^2 \\ (w_{ji}(\Sigma_G)_{jj})^T_{ij} \\ \vdots \\ (w_{ji}(\Sigma_G)_{jj})^T_{ij} \end{pmatrix}.$$
Conditional Distribution of $G$ given $Y$

**Lemma 5.A.1.** Let $G \sim \mathcal{N}_p(\mu, \Sigma)$ and $Y$ be defined as

$$Y_j \mid G_1, \ldots, G_p \sim \mathcal{N}(\sum_{k=1}^p w_{kj} G_k, \sigma^2) \quad (j = 1, \ldots, p).$$

Then $(G \mid Y)$ is normally distributed with expectation

$$E[G \mid y] = (K + \frac{1}{\sigma^2} WW^T)^{-1}(K\mu + \frac{1}{\sigma^2} Wy),$$

where $K = \Sigma^{-1}_G$ and the matrix $W = (w_{ij})_{i,j=1,\ldots,p}$.

**Proof:**

The Bayes’ theorem gives us

$$f(g \mid y) = \frac{f(y \mid g)f(g)}{\int f(y \mid g)f(g)dg}.$$ 

The numerator can be rewritten by using the known distributions of $G$ and $Y \mid G$. 

$$\log(f(g)f(y \mid g)) \propto -\frac{1}{2}(g - \mu)^T \Sigma^{-1}_G (g - \mu) - \frac{1}{2\sigma^2}(y - W^T g)^T(y - W^T g)$$

$$= -\frac{1}{2}g^T(K + \frac{1}{\sigma^2} WW^T)g + g^T(K\mu + \frac{1}{\sigma^2} Wy)$$

$$- \frac{1}{2}\mu^T K\mu - \frac{1}{2\sigma^2} y^T y$$

We define the following three quantities

$$M = K + \frac{1}{\sigma^2} WW^T, \phi = K\mu + \frac{1}{\sigma^2} Wy \quad \text{and} \quad a = -\frac{1}{2}\mu^T K\mu - \frac{1}{2\sigma^2} y^T y.$$ 

Then, the equation above simplifies to

$$\log(f(g)f(y \mid g)) \propto -\frac{1}{2}g^T Mg + g^T \phi + a.$$
We add an additional term to get a quadratic form:

\[
\log(f(g)f(y | g)) \propto -\frac{1}{2} g^T M g + g^T \phi - \frac{1}{2} \phi^T M^{-1} \phi + \frac{1}{2} \phi^T M^{-1} \phi + a
\]

\[
= -\frac{1}{2} (g - M^{-1} \phi)^T M (g - M^{-1} \phi) + c_1
\]

where

\[
c_1 = \frac{1}{2} \phi^T M^{-1} \phi + a.
\]

Note \(c_1\) is independent of \(g\).

Let \(c_2 = (f(y))^{-1}\). Then the conditional density of \(G\) given \(Y\) is equal to

\[
f(g | y) = c_2 \exp(c_1) \exp(-\frac{1}{2} (g - M^{-1} \phi)^T M (g - M^{-1} \phi)).
\]

Hence the expectation of \(f(g | y)\) is

\[
M^{-1} \phi = (K + \frac{1}{\sigma^2} W W^T)^{-1} (K \mu + \frac{1}{\sigma^2} W y).
\]

\[\square\]

## 5.B Reformulation as Lasso Problem

The proof of the following theorem is based on Friedman et al. 2007.

**Theorem 5.B.1.** The solution that is obtained from the convex optimization problem

\[
\hat{K} = \arg \max_{K > 0} \left( \log |K| - \text{tr}(K S) - \rho \|K\|_1 \right)
\]

for the last column or row is equivalent to

\[
\hat{\beta} = \arg \min_{\beta} \left( \frac{1}{2} \|V_{11}^{1/2} \beta - V_{11}^{-1/2} s_{12}\|^2 + \rho \|\beta\|_1 \right)
\]

\[
v_{12} = \frac{1}{2} V_{11} \hat{\beta} \quad \text{and} \quad v_{22} = s_{22} + \rho
\]

where \(V\) is the estimate of \(\Sigma = K^{-1}\).
**Proof:**
Define the function which we want to optimize as
\[
    f_1(K) := \log|K| - \text{tr}(KS) - \rho\|K\|_1 \tag{5.4}
\]
respectively,
\[
    f_2(\beta) := \frac{1}{2}\|V_1^{1/2}\beta - V_1^{-1/2}s_{12}\|^2 + \rho\|\beta\|_1. \tag{5.5}
\]
The sub-gradient equation of (5.4) can be calculated as
\[
    \nabla f_1(K) = K^{-1} - S - \rho\Gamma_K = \Sigma - S - \rho\Gamma_K
\]
where
\[
    \Gamma_K = (\gamma_{ij}) = \begin{cases} 
        \text{sign}(k_{ij}) & \text{if } k_{ij} \neq 0 \\
        \in [-1, 1] & \text{if } k_{ij} = 0.
    \end{cases}
\]
Reformulating this for the last column of \(\Sigma = K^{-1}\) leads to
\[
    \nabla f_1(\sigma_{12}) = \sigma_{12} - s_{12} - \rho\gamma_{12}
\]
and
\[
    \frac{\partial}{\partial \sigma_{22}} f_1(\sigma_{22}) = \sigma_{22} - s_{22} - \rho\gamma_{22}
\]
If we set this equation to zero, we retrieve the estimates \(v_{12}\) and \(v_{22}\)
\[
    v_{12} = \hat{\sigma}_{12} = s_{12} + \rho\gamma_{12}
\]
\[
    v_{22} = \hat{\sigma}_{22} = s_{22} + \rho \quad (\text{since } \sigma_{22} > 0).
\]
The function (5.5) can be rewritten as follows
\[
    f_2(\beta) = \frac{1}{2}\|V_1^{1/2}\beta - V_1^{-1/2}s_{12}\|^2 + \rho\|\beta\|_1
\]
\[
    = \frac{1}{2} (\beta^T V_1 \beta - 2\beta^T s_{12} + s_{12}^T V_1 s_{12}) + \rho\|\beta\|_1.
\]
Therefore, the sub-gradient equation is equal to
\[
    \nabla f_2(\beta) = V_1^{1/2} - s_{12} + \rho\Gamma_\beta.
\]
We substitute \(\beta := V_1^{-1}v_{12}\). This yields
\[
    \nabla f_2(\beta) = v_{12} - s_{12} + \rho\Gamma_{(V_1^{-1}v_{12})}.
\]
Using the fact that $V K = I$ and $k_{22} > 0$, we have

\[ V_{11} k_{12} + v_{12} k_{22} = 0. \]

Therefore,

\[ \text{sign}(V^{-1}_{11} v_{12}) = -\text{sign}(k_{12}) = -\gamma_{12} \]

and

\[ \nabla f_2(\beta) = v_{12} - s_{12} - \rho \gamma_{12} = \nabla f_1(\sigma_{12}). \]

\[ \square \]

5.C The EM Algorithm

The Expectation-Maximization (EM) algorithm is an iterative optimization method in the case of incomplete data. Let $Y$ denote the observed data and $G$ denote the unobserved latent data. We define the complete data as

\[ T = (Y, G). \]

We want to estimate the parameter $\Theta$ by optimizing the observed log-likelihood $\log(f_Y(y; \Theta))$. The EM algorithm finds an estimate for the complete likelihood function $\log f_T(t; \Theta)$ and then maximizes it for the parameter $\Theta$.

**EM Algorithm**

1. Set $m = 0$ and initialize $\hat{\Theta}^{(0)}$.
2. Iterate 3 - 6 until convergence.
3. **Iteration Step** $m + 1$:
   4. $q(\Theta | \hat{\Theta}^{(m+1)}) = E[\log f_T(t; \Theta) | Y, \hat{\Theta}^{(m+1)}]$ \hspace{1cm} (E-Step)
   5. $\hat{\Theta}^{(m+1)} = \arg \max_\Theta q(\Theta | \hat{\Theta}^{(m+1)})$ \hspace{1cm} (M-Step)
4. Set $m = m + 1$

The following theorem by Dempster et al. 1977 shows that the EM algorithm increases the observed log-likelihood $\log(f_Y(y; \Theta))$ in every iteration step:
Theorem 5.C.1. For EM algorithm holds
\[ \log(f(y; \hat{\Theta}^{(m+1)})) \geq \log(f(y; \hat{\Theta}^{(m)})) \]
i.e. the MLE improves in each step.

Proof:

We can write
\[ \log(f_Y(y; \Theta)) = E[\log(f_T(y, g; \Theta)) \mid Y] + E[\log(f_{G \mid Y}(g\mid y; \Theta)) \mid Y] \]
since by the Bayes’ rule holds
\[ f_Y(y; \Theta) = f_T(y, g; \Theta) / f_{G \mid Y}(g \mid y; \Theta). \]

We add an additional term to the equation in order to obtain a term which is equal to the the Kullback-Leibler Divergence \( KL \):
\[ \log(f_Y(y; \Theta)) = E[\log(f_T(y, g; \Theta)) \mid Y] + E[\log(f_{G \mid Y}(g\mid y; \Theta^{(m)})) \mid Y] \]
\[ - E[\log(f_{G \mid Y}(g\mid y; \Theta)) \mid Y] + E[\log(f_{G \mid Y}(g\mid y; \Theta)) \mid Y] \]
\[ = E[\log(f_T(y, g; \Theta)) \mid Y] + E[\log(f_{G \mid Y}(g\mid y; \Theta^{(m)})) \mid Y] \]
\[ - KL(\Theta, \Theta^{(m)}) \]

We need to show that the following subtraction is greater than zero.
\[ \log(f(y; \hat{\Theta}^{(m+1)})) - \log(f(y; \hat{\Theta}^{(m)})) \]
\[ = E[\log(f_T(y, g; \Theta^{(m+1)})) \mid Y] - E[\log(f_T(y, g; \Theta^{(m)})) \mid Y] \]
\[ - KL(\Theta^{(m+1)}, \Theta^{(m)}) + KL(\Theta^{(m)}, \Theta^{(m)}) \]

For the Kullback Leibler Divergence holds \( KL(f_1, f_1) = 0 \) and \( KL(f_1, f_2) > 0 \) for any densities \( f_1 \) and \( f_2 \). Therefore,
\[ \log(f(y; \hat{\Theta}^{(m+1)})) - \log(f(y; \hat{\Theta}^{(m)})) \]
\[ \geq E[\log(f_T(y, g; \Theta^{(m+1)})) \mid Y] - E[\log(f_T(y, g; \Theta^{(m)})) \mid Y] \]

In the M-Step, exactly these quantities are maximized. Therefore, we have
\[ E[\log(f_T(y, g; \Theta^{(m+1)})) \mid Y] \geq E[\log(f_T(y, g; \Theta^{(m)})) \mid Y]. \]
It follows
\[
\log(f(y; \hat{\Theta}^{(m+1)})) - \log(f(y; \hat{\Theta}^{(m)})) \geq 0
\]
which completes the proof.

Additional conditions are needed that EM converges to a local maximum of log likelihood but it is not clear if a global or local maxima is the best fit.
Conclusions and Outlook

We proposed several statistical methods for studying large-scale data coming from RNAi experiments and yielded novel results on the biological process of infection by applying the methods to the screening data of the InfectX consortium. In the following, we list all the main findings of this thesis and give an outlook on open questions that could be addressed in future.

6.1 Theoretical Findings

We proposed the PMM which simultaneously analyzes the data of parallel RNAi screens and enriches each hit with an FDR estimate (Rämö, P and Drewek, A et al. 2014). PMM improves the reliability of hit detection in individual RNAi screens through its higher statistical power compared to commonly used methods. Even a slight overlap between the studied biological conditions can improve the hit detection in individual screens without compromising the detection of unique hits. Another beneficial feature of PMM is the possibility
to assign weights to the individual data points. The weights can, for example, incorporate a-priori information about the performance of individual siRNAs or the quality of the phenotypic readouts. PMM could be further developed in order to account for dependencies between the genes. Moreover, it could be interesting from biological side to add an additional dimension to the PMM for the analysis of more than one experimental condition.

A major contribution of this thesis is the availability of the PMM code as R package (Drewek 2015). This allows other scientists to fit the PMM to their RNAi data and to benefit from the gain of statistical power. PMM can, in principle, be applied to any kind of parallel RNAi screens almost independently of the underlying biology or field of application, as long as the readouts of the screens are measured on the same scale.

In the second part of this thesis we proposed a set of statistical methods to analyze genome-wide RNAi screens having only few replicates. We showed that the reliability of hit detection can be improved by involving results from miRNA mimic screens. Our proposed model quantifies the marginal readout effect caused by a single gene from the readout of miRNA mimic screens. Therefore, it enables a direct comparison between miRNA mimic screens and siRNA screens. Furthermore, we proposed an adaptive thresholding technique to achieve a double checked hit list from a full primary RNAi screen and a ”cherry-picked” secondary RNAi validation screen using the hypergeometric distribution.

The off-target correction of RNAi screening data is a very recent and challenging topic. We contributed to this with the refinement of an existing method. However, the question of the goodness of our refined model was not answered within this thesis, as it has not been answered for all the other suggested methods. The reason is that at the moment there are no suitable instruments available to prove hypothesis on off-targets from the experimental side. The new technique CRISPR/Cas9 (Cong et al. 2013) seems promising to solve this problem as soon as it is available for high-throughput screening.
In the last chapter, we showed an algorithm that exploits the off-target effects as multiple knockdowns and reconstructs from them an estimate for the inverse covariance matrix of the pure gene effects. The EM algorithm was able to recover the inverse covariance matrix in our simulation settings. Further work is needed to derive theoretical results on the consistency of the estimator. In particular, the irrepresentability condition needs to be checked for our underlying model.

6.2 Applicational Findings

We applied our proposed statistical methods to the RNAi screens of the InfectX consortium. Adding up the detected hits from the kinome-wide pathogen screens and the genome-wide screens, we detected around 264 new interesting hits which have an influence on the infection. These hits require further follow-up experiments in order to understand the exact biological mechanism of the each detected gene and to remove false negatives. If some of the genes are confirmed and turn out as independent from essential survival processes of the human cell, researches could go one step ahead and test the knockdown of these genes as potential blockade to pathogens. There is a high probability that this set of genes has an increased therapeutical effect since 30% of our detected hits are shared between the pathogens.

The RNAi screening data of the InfectX consortium were generated through highly unified protocols and common image analysis. This data analysis pipeline enabled a direct comparison between the phenotypic readouts of the different pathogen screens. This also applied to the feature data of the individual cells. For our analysis we used the 300 standard features. The individual cell data have definitely a so far unused potential. One could think of developing more specific features in collaboration with the biologists in order to have a more profound understanding of the individual cell process.


