Statistical analysis tools for metabolic and genomic bacterial data

Minna Vehkala

Academic dissertation to be presented, with the permission of the Faculty of Science of the University of Helsinki, for public examination in Exactum, Auditorium CK112, on December 2nd, 2016, at 2 o’clock in the afternoon.
Abstract

This thesis introduces statistical analysis methods for two types of bacterial data: metabolic data produced by phenotype microarray technology, and genomic data produced by sequencing technologies. As both technologies produce vast amounts of data, as well as have special features, there is a need for bioinformatics tools that adequately process and analyze the information produced. Similar to all biomolecular data analyses, the interplay between biological components poses an additional challenge to the method development. A specific complication, regarding the metabolic data, is the lack of larger quantities of replicates due to the high expenses of performing the experiments. In terms of the sequence data, genome-wide analysis tools are desired, since such methods have not yet been widely developed for bacteria, even though they exist for eukaryotic genetics. The thesis briefly reviews the current methods, and introduces new approaches tackling the above mentioned problems.

General Terms:
biomolecular data, bioinformatics, method development, bacteria

Additional Key Words and Phrases:
Biolog phenotype microarray data, metabolic activity, biochemical substrate, time-series, multidimensionality, dimension-reduction techniques, FLOG, DNA, genome, genotype, phenotype, genome-wide association study (GWAS), k-mer, SEER, logistic/linear regression analysis, multiple hypothesis testing, confounding effect, population structure
Acknowledgements

After a journey one is always happy to be back at home, no matter how nice and wonderful the trip has been. Similarly, I am happy, and also extremely relieved, to end the journey of becoming a PhD. Particularly happy I am to end it to a new home, the building of which has been the other huge project during the past year in addition to completing the PhD.

On the way, I have been entitled to get to know many amazingly intelligent people showing the greatest passion and devotion towards research. Especially, the enthusiasm of my supervisor Jukka Corander towards science is impressive. Such overwhelming positive energy is rarely present in a person. Most wonderfully the topics of interest he has are not only limited to science: any matter whether related to sports, travelling or house building can be discussed with Jukka. I want to thank Jukka for giving me the opportunity to work my way to a PhD. I appreciate the freedom in the working schedules and being able to work from home whenever needed. This has enabled me to enjoy also other aspects of life, such as joining my husband when travelling to Pohjanmaa for farming duties and arranging spectacular holidays. During these years I have made the most unbelievable trips with my husband and friends to all over the world. Of course, the many scientific conferences, I have been entitled to attend to, have been wonderful experiences as well, especially the latest trip to Japan.

Since I started as a PhD student, the impressively large group of PhD students of Jukka has become smaller and smaller, as one by one we have got our PhDs. I am among the last ones to finish, but I have had the privilege to learn about the process of becoming a PhD by watching my colleagues. Thus, I would like to thank you guys for helping me (maybe even without knowing) to prepare myself for the PhD and the foremost the defence. I would also like to thank you for your efforts in organizing and attending the Statistics discussion club, lunches and evening get-togethers. As we all know, the academic work itself often is lonely and exhaustive, but lunch breaks, especially with Elina and Mikhail, and discussions with Jukka K. and Paul sitting in the same office made the days more enjoyable.
Workwise, I am the most grateful to Mikhail for his collaboration. I highly appreciate his visual and technical inputs to the Biolog projects as well as the critical inspection of our articles.

I would also wish to express my gratitude to Siru Varvio for her support in organizing me teaching activities on many statistics courses. These duties provided me the much needed breaks from doing research, widened my perspectives on the administrative side of the university, and let me work and interact with younger students. I value the enormous amounts of work Siru does for students, both younger and older, and hope her efforts are appreciated by the faculty as well.

At the very final steps three people, unknown to me, had a significant role in making the PhD reality. Namely, it would have not been possible to complete the PhD without the efforts made by the pre-examiners Tanel Tenson and Jing Tang and the most of all the opponent Zhaohui Qin. I am greatly thankful for their efforts.

To my husband I must say: “You are my pride and joy, the most precious thing in my life and the touch to reality. Therefore I am truly sorry for making you tolerate the crabby side effects of this work: all the frustration, stress and anxiety. If I had even a small portion of the energy and enthusiasm by which you have built us two houses, this project would have been completed ages ago. While waiting for me to get a real job (and start earning money), thank you for your endless patience towards my Savonian (slow, modest, passive) temper, and the most of all for providing us a good standard of living including many relaxing trips, an amazing wedding and gourmet dinners.”

Finally, I want to dedicate this book to my grandmother for showing a constant interest towards my work, even though most of the times we meet in Kuopio, I am in holiday mode which does not make me eager to talk about work related topics. However, I believe she is the person most proud of this achievement.
List of articles


II Minna Vehkala, Mikhail Shubin, Jukka Corander. (2016) FLOG: Multivariate R Tool for Exploring Biolog Phenotype Microarray Data, Submitted Manuscript


**Author contributions**

**Articles I and II**: MV had the main responsibility in designing and implementing the analysis methods as well as in writing the articles. MS contributed in study design, implementation and writing the articles. JC participated in method development and editing the articles. TRC and NRT provided data and biological expertise.

**Article III**: MV had the main responsibility in implementing the initial version of the method designed by JC. JL had the main responsibility in the implementation of the final version as well as in writing the article, while MV participated in writing. NV, PM, AH took part in method design. SRH, CC, NJC, MRD, ACS, SYCT, JP and SDB provided data and biological expertise.

**Article IV**: MV participated in analyzing the data jointly with NV, LW, SP, TJ, RC, JC, AB and KH. DM, AT, AR, PL, BW, SB, MH, JP, NTH and JF conceived the study. JW, SP, DH, TTBC, NVVC, JC, NTH and CS produced the data. LW, DM, AT, MH, SP, JC, DH and AB wrote the article.
## Contents

1 Introduction ........................................ 1

2 Biolog Phenotype Microarray (PM) technology ........ 5
   2.1 Analysis methods for PMs ....................... 7
   2.1.1 Methods introduced in article I ............ 12
   2.1.2 Methods introduced in article II .......... 16

3 Genome-wide association studies (GWAS) ............. 19
   3.1 GWAS in bacteria .................................. 22
   3.1.1 Sequence element enrichment analysis (SEER) 24
   3.1.2 Applications of SEER to real data .......... 28

4 Discussion ........................................... 35

References ............................................ 37
Chapter 1

Introduction

The revolution of measurement technology has also revolutionized the study of cells on molecular level within the past decades, for example, the human genome was first sequenced in 2001 [1,2]. Back then it required 13 years to complete the work. After that basal work the developments to the sequencing machines have been enormous, reducing the time spent to sequence a human genome first to days, and recently to hours. The advances in the sequencing technology have led to substantial reductions in the costs as well, allowing sequencing to become a standard procedure in health care used for clinical diagnostic testing [3,4]. Sequencing techniques are not the only measurement techniques improved, other examples include DNA microarrays, mass spectrometry, and cellular imaging by microscopy [5]. One such measurement technique is Biolog phenotype microarray technology elaborated in this thesis [6].

The biomolecular measurement techniques of this era produce vast amounts of digital data, creating the need for computational tools to effectively process the data and to handle both biological and technical noise embedded in the observations. In this thesis, we focus on developing such analysis tools for Biolog phenotype microarray data as well as for DNA sequence data. Especially, we concentrate on detecting similarities or differences in either the metabolic or genomic composition of bacterial samples. The statistical R programming environment is mainly utilized to accomplish the tasks.

Biolog phenotype microarrays measure the metabolic activity of cells in thousands of predetermined conditions, and are mainly applied to bacteria [7–9], even though the metabolic activity of fungi [10], yeast [11], human [12] or virus infected cells [13] can be studied as well. Not many publicly available tools exist for processing, exploring and analyzing the data which by nature are multidimensional, as thousands of phenotypes
can be monitored at once, and followed over a period of time. Thus, we developed a pipeline from preprocessing to the detection of differentially metabolized samples. The former includes steps such as background correction and normalization, and the latter can be achieved, for example, by variance or factor analysis. The approaches we have developed are described in articles I and II.

The DNA sequence data addressed in this thesis are bacterial as well, and the aim is to track genetic variants associated with a phenotype of interest, such as antibiotic, or multi-drug resistance, host, and geographical location. Such associations between the genetic composition and an outcome variable are traditionally studied by linkage or genome-wide association analyses (GWAS) [14,15], originally developed to study human genomes [16,17], but recently also applied to bacteria [18–20]. However, as human and bacterial cells differ from each other, e.g., in terms of gene content, recombination rate and clonality, the methods developed for human lack power to detect associations in bacteria. Our method, called sequence element enrichment analysis (SEER), tackles the problems of variable gene content and clonality by using sequence elements, i.e. DNA words of length \( k \), hence spanning the search of genetic variants to cover the whole genome, instead of only focusing on the core genome parts, which is done in the traditional approaches utilizing SNPs (single nucleotide polymorphism). The non-core genes do often contribute substantially to several phenotypes of interest, such as antibiotic resistance, and consequently SNPs in core genes are not necessarily able to discriminate between different phenotype values [21]. So far, SEER has been applied to tens of studies, providing numerous new insights into the relationships between genotypes and phenotypes [21–23]. Article III represents the SEER method, while article IV introduces an early version of SEER and includes the first ever application of the SEER approach.

This thesis is divided into two parts: Chapter 2 introduces the Biolog phenotype microarray technology with existing analysis methods as well as the tools developed here, whereas Chapter 3 focuses on the genome-wide association analysis of DNA sequence data. Finally, Chapter 4 concludes the thesis with a discussion. Throughout the thesis, some technical concepts are highlighted by blue colour, and briefly explained after each section.
**Molecular biology** = a science that studies the composition, structure and interactions of cellular molecules, such as DNA, RNA, and proteins, to understand the complex biological processes vital to cell maintenance.

**Bioinformatics** = Molecular biology + Statistics + Computer science

**Metabolic activity** = degree of activity of the process called metabolism through which a cell gets energy from nutrients to synthesize new proteins, nucleic acids (DNA, RNA), etc.

**DNA sequence** = a string of nucleotides constructing the genetic material of each living cell

**Genetic variant** = a section of DNA sequence that differs within or between populations

**Recombination** = exchange of genetic material between chromosomes (eukaryotes) or cells (prokaryotes)

**Clonality** = an offspring inherits the genes of its parent, resulting in an offspring identical, or nearly so, to its parents

**SNP** = a single nucleotide difference at a certain position of the DNA sequence

**Whole genome** = all the genetic material of an organism

**Core genome** = genetic material present in all the genomes compared

**Non-core gene** = a gene not present in all the genomes compared
1 Introduction
Chapter 2

Biolog Phenotype Microarray (PM) technology

Biolog phenotype microarray (PM) technology is a high-throughput technique developed at the beginning of the 21st century allowing simultaneous testing of thousands of phenotypes that represent a significant fraction of the functions cells can perform [6, 24, 25]. Similar to more common microarray platforms, such as Affymetrix or Illumina, PMs are colorimetric assays, but instead of gene expression, PMs measure the ability of a cell line to metabolize a biochemical substrate, and hence to produce energy. If energy is produced, i.e. cell respiration occurs, it is detected as an irreversible colour change of a redox dye.

Unlike gene expression microarrays that use RNA as an input, PMs monitor substrate metabolism of living cells. Also, instead of making a single measurement at a prefixed time point, the PM experiment is kinetic, run over a period of time, usually for 24 or 48 hours, allowing real time monitoring of the flow of energy production and cell respiration.

As shown in Figure 2.1, when performing a PM experiment, cells are directly pipetted into 96-well plastic microplates, making a PM experiment more straightforward to perform than a gene expression study that requires extraction and reverse transcription of RNA. On the microplates, the 96 wells introduce 96 different preconfigured tests. In both methods electronic scanners are used for detecting colour changes. In contrast to RNA microarrays for which visual inspection is not possible, the results of the PM approach can be also visually verified as colouring is detectable by eye as demonstrated in Figure 2.1

Each well on a PM plate contains a different substrate, such as carbon, nitrogen, phosphorus or sulphur source, hormone, anti-cancer agent or antibiotic, some at varying concentrations. These metabolic substrates on
PM panels can be linked to cellular pathways using public databases such as KEGG [26].

In some wells, the cells are activated by the given substrate, simultaneously reducing the redox dye and forming purple colour, whereas in some wells the cells are inhibited with little or no colour formed. The intensity of the purple colour is measured and recorded by the accompanied OmniLog™ incubator-reader, usually every 15 minutes. In each well, the amount of purple colour reflects the amount of cell respiration. As cell respiration can occur independent of cell growth, PM technology allows measuring phenotypes that do not necessarily lead to growth. To identify differences in substrate metabolism, 20 preconfigured PM panels are available for microbial and 14 for mammalian cells, yielding the capacity of testing nearly 2,000 and 1,400 different response phenotypes in each category, respectively. Since the OmniLog™ incubator-reader contains 50 plate holders, as illustrated in Figure 2.1, almost 5,000 wells can be simultaneously monitored.

PMs were originally developed for microbial cells (bacteria, yeast and fungi) to analyze the effects of loss of gene function [6], but have later been established also for human cells as understanding changes in substrate metabolism is important in diseases, such as diabetes, obesity, and cancer [27]. PM panels for human provide carbon and nitrogen substrates as well as sensitivity tests against ions, hormones, cytokines and well-established anti-cancer chemical agents. The metabolic profiles of diseased
2.1 Analysis methods for PMs

Even though the PM technology has been available for almost two decades, the concept is not as widely exploited as many other high-throughput methods, such as gene expression microarrays. The low utilization rate is partly due to the high costs of setting up a Biolog testing environment as the OmniLog™ incubator-reader is an expensive investment not all laboratories can afford to, and partly due to the lack of knowledge in analysing the data. As few scientists are running Biolog experiments the development of analysis methods has not been especially intensive either, and the majority of the Biolog data analyses are performed by plotting two metabolic profiles on the top of each other after which differential metabolism is detected either by visual inspection or based on arbitrary cut-off values [6, 24, 25]. Additionally, the experiments are often performed on a limited number of samples and replicates, thus hindering a proper analysis of the resulting data and complicating the interpretation of the results.

As demonstrated in Figure 2.2, the software included in the OmniLog™ system allows the comparison of two strains based on a 96-panel chart with...
Biolog Phenotype Microarray (PM) technology graphically overlaid metabolic profiles. The amount of purple colour formed in each of the 96 wells throughout the time course of an experiment is represented as curves in the 96 panels. When comparing two metabolic profiles, one profile is shown by green, the other by red, and the overlapping area by yellow colour. Additionally, the software calculates parameters from the kinetic data, such as the minimum, maximum and average response, area under the curve (AUC), the length of the lag phase and slope. Any of the above-mentioned kinetic parameters can be utilized to highlight the wells in which the metabolic activity between the two profiles differs more than a given threshold. Clearly, no statistically sound inference is made when comparing two curves without replicates, and using an arbitrary threshold. Another disadvantage, related to using a single summary statistic to describe a curve which originally comprises hundreds of data points, is loss of information concerning the shape of the underlying curve. For instance, two curves may have very similar average response and AUC, but yet clearly differ in their shapes, for example, in terms of the length of lag phase, maximal growth rate and maximum response [34].

In a Biolog experiment, the colour accumulates irreversibly, thus the recorded metabolic profiles are increasing curves, enabling the use of growth models in the analysis of PM data. Several models have been suggested to fit the Biolog metabolic growth curves, such as logistic, Gompertz, Lindstrom, Richard, Baranyi, and Diauxic [11,34–37]. After fitting a model, the curves are compared in a similar fashion as above, but using the model-based parameters instead of the non-model-based summaries, the benefit being that the standard errors of the estimated parameters can be utilized in assessing statistically significant differences. The disadvantage of the model-based approach is that there exists no single model that could depict the variety of curve shapes introduced by Biolog experiments [34,37]. Ideally, the same model should be fitted to all metabolic curves to be reliably able to compare the resulting parameter estimates, since the parameters of different models might not be comparable. Gerstgrasser et al. [34] solves this problem by fitting several models, and using summary statistics calculated on the basis of the fitted curves instead of the model parameters.

Probably the most widely used and comprehensive package for reading in, processing, and visualizing PM data was introduced by Vaas et al. [37]. The R package is called opm, and it fits Gompertz’s and Richard’s models to the Biolog metabolic activity profiles, and also provides a model-free spline fit. Differences between the parameters of different curves can be evaluated based on the 95% confidence intervals of the estimated parameters.

As stated above, Biolog experiments are often performed with no or few
Figure 2.2: Comparison of two samples. A Biolog phenotype microarray experiment is performed to compare the metabolic activity of samples A and B. Left: Two 96-well PM plates are loaded with cells from two different cell lines. Purple colour starts to develop if the substrate in a well triggers metabolism. The darker the colour is, the more metabolically active the cells are. Middle: The plates are placed in the OmniLog™ incubator-reader which tracks the amount of purple colour produced in the wells for a period of time. Top right: The output from the scanner is time-series data indicative of the metabolic activity of the cells under the conditions provided in each well. The metabolic activities of the samples A and B are represented by red and green colours, respectively. The overlapping section is coloured by yellow. Bottom right: Several summary statistics can be extracted from the metabolic activity curves, such as length of lag phase, maximum growth rate, area under the curve (AUC), or minimum and maximum signal.
replicates complicating the statistical analyses. However, if an experiment is repeated several times, statistical tests, e.g. t-test (for two groups) or ANOVA (for several groups) can be applied to the parameters summarizing the profiles [38]. If several samples are available, the parameter vectors can be collected into a matrix as illustrated in Figure 2.3, a covariance or distance matrix computed, and used as an input, for instance, for factor analysis (FA), principal component analysis (PCA), multidimensional scaling (MDS) [35] or hierarchical clustering. The described approaches aim to identify groupings within samples. Alternatively, groupings within substrates, can be identified by calculating the pairwise covariances or distances between the substrates rather than between samples. Another common method in addition to using summary statistics or model parameters, is to make a binary growth versus no-growth distinction, for example, based on an arbitrary threshold or by comparing curves to a reference curve [7,31]. In this approach, the distances between the resulted binary vectors can be defined as Hamming distances. One recently published method describes individual metabolic curves in terms of the number and nature of metabolic cycles present in an experiment due to sequential use of different metabolic pathways, or the presence of subpopulations in the samples [39]. Similar to any other summary statistic, these measures can be utilized when comparing samples.

Similar to gene expression microarray data, preprocessing of PM data is required before performing any of the above described statistical analyses. The most essential preprocessing steps, background correction and normalization, make the samples comparable with each other, and thus the results more reliable. All PM plates contain at least one control well (usually well A01) and in the background correction, the signals produced by the control well are subtracted from the other metabolic profiles. Normalization, on the other hand, removes systematic errors from the experimental data. Such errors may be caused by differences, for example, in array quality, sample preparation, equipment, laboratories, technicians, or number of cells pipetted into the plates.
2.1 Analysis methods for PMs

Figure 2.3: A workflow for detecting similar metabolic activity patterns within samples. First panel: raw or background corrected metabolic curves of $n$ samples. Second panel: summary statistics of $n \times 96$ curves collected into a matrix. Third panel: pairwise distances between samples, for example, based on Euclidean distance or covariance. Fourth panel: clustering of samples by dimension reduction methods, such as FA, PCA and MDS, or hierarchical clustering.
Area under the curve (AUC) = the area under the metabolic activity curve

Lag phase = the phase prior to rapid growth. During the lag phase, bacteria adapt themselves to the substrate conditions.

Slope = describes the steepness of the curve/maximal growth rate

Growth model = a mathematical model that can simulate a process studied over time

Spline fit = a piecewise polynomial function where each sub-function is most commonly a cubic function

Factor analysis (FA) = a statistical procedure that converts correlated variables into a smaller set of linearly uncorrelated variables called factors. The observed variables are modelled as linear combinations of the factors, plus error terms, i.e. $X = CF + E$, where $X$ is a matrix including the observed variables, $C$ a matrix of loadings, $F$ a matrix including the factors, and $E$ a matrix of error terms.

Principal component analysis (PCA) = a statistical procedure that converts correlated variables into a smaller set of linearly uncorrelated variables called principal components. The principal components are modelled as linear combinations of the observed variables, i.e. $P = CX$, where $P$ is a matrix including the principal components, $C$ a matrix of loadings, and $X$ a matrix including the observed variables.

Multidimensional scaling (MDS) = a statistical procedure that projects observed distances of study objects (e.g. samples, or substrates) into a reduced number of variables/dimensions by minimizing the change in the between-object distances. In contrast to PCA and FA, any similarity or dissimilarity matrix, in addition to correlation matrix, can be used.

Hierarchical clustering = a statistical procedure that converts the distances between the objects studied into a dendrogram. For example, Euclidean or Hamming distances can be used as a dissimilarity measure between objects.

Euclidean distance = in a two-dimensional space the distance between points $p = (p_1, p_2)$ and $q = (q_1, q_2)$ is $\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$. For n-dimensional space, this can be extended as $\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$.

Hamming distance = for two vectors of equal length, the number of positions at which the corresponding symbols are different

2.1.1 Methods introduced in article I

In article I, we introduce a three-step pipeline for analyzing PM data including 1) a binary grouping of the metabolic curves into active and inactive, 2) normalization, and 3) comparison of samples. All the methods are motivated by the hypothesis that the metabolically active and inactive wells should be treated separately when analyzing PM data. Especially, the existing methods for normalizing PM data suffer from the combined analysis of the inactive and active wells, and in the comparison of samples, the inclusion of wells in which no metabolic activity occurs, is not necessary either.
2.1 Analysis methods for PMs

Figure 2.4: Grouping of metabolic activity curves into active and inactive by using the EM-algorithm. The 96 growth curves shown by pink and yellow colours represent the metabolic activity curves produced by the 96 wells of a single plate. Black dotted line at 100 is a cut-off for the activity (curves not exceeding this level are never considered as active). At each iteration, two steps are performed: 1) grouping and 2) model fitting. Iteration 1: 1) an initial grouping into active and inactive is performed based on the activity cut-off, i.e. curves not exceeding the cut-off of 100 at any time point are assigned to the inactive group (yellow), whereas the curves exceeding the cut-off at least at one time point are assigned to the active group (pink), then 2) a mixture of linear and logistic model (linear to the inactive and logistic to the active) is fitted, and shown as black solid lines. Iteration 2 & 3: 1) each curve is reassigned either to the active or inactive group based on whether it resembles more the linear or the logistic curve fitted at the preceding iteration, then 2) the mixture model is refitted according to the new grouping. Here, only three iterations are needed for the solution to converge.

The grouping into active and inactive is done by applying the Expectation-Maximization (EM) algorithm [40, 41] iterating between two states: 1) fit a mixture of a linear and logistic model, and 2) assign each curve to the most probable group. The linear model is reserved for the inactive, whereas the s-shape of the logistic curve is assumed to describe well the curves produced by the metabolically active wells. The EM algorithm proceeds until the likelihood of the mixture model convergences to its maximum, or the maximum number of iterations is reached. All the curves on a plate are evaluated at once, resulting in a grouping that is relative to the degree of metabolic activity on a plate, i.e. if the metabolic activity on a plate is in general low, curves with a relatively low metabolic activity are addressed as active. The comparability between plates can be enhanced by defining a cut-off for the activity, i.e. the curves not exceeding the given value are never labelled as active. The same cut-off is utilized to define the initial grouping, which in part helps to restrict the number of iterations. Figure 2.4 illustrates the step-wise EM approach for grouping.

The existing solutions for normalization can result in biased curves, as they often divide the raw signals by a measure called the average well colour development (AWCD) which becomes under-estimated if many of the curves are inactive. Additionally, the resulting normalized curves often
lack the shape of growth curves, preventing the fitting of growth models to them. Our normalization method utilizes the above gained distinction into active and inactive curves as well as considers the fitted mixture model as the average metabolic behaviour on a plate. However, as the average metabolic profiles of replicated measurements are compared and adjusted against each other, at least three replicates are required to accomplish the normalization.

The third step, comparison of samples, utilizes the grouping into active and inactive as well. Now, linear or logistic model is fitted to a single curve instead of a group of curves. In comparison to the existing methods [37], the aim of model fitting is noise reduction rather than comparison of the resulting parameter estimates. Hence, instead of using the model parameters, we use the predicted values to detect differences in metabolic activity between samples. The statistical testing is performed by applying a Bayesian two-way analysis of variance (ANOVA) at several time points. The approach enables testing the main effects of two categorical independent variables on the metabolic activity as well as their interaction effect. In the experimental data represented in the article, the metabolic activity of several *Yersinia enterocolitica* strains is observed at different temperatures. In this example, we are able to assess the main effects of strain and temperature and their interaction. In the case of two strains tested at two temperatures, interaction can, for example, imply to strains differing in their metabolic activity at the higher temperature, but showing no differences at the lower temperature. Such a pattern is illustrated in Figure 2.5.

---

1Temperature can be easily adjusted by the OmniLog™ incubator-reader, and for many bacteria their metabolic activity at different temperatures is of interest.
Figure 2.5: Detection of differences in the metabolic activity samples. Left: 12 normalized metabolic activity curves are categorized into four groups according to temperature and strain, each setting including three replicates. Coloured dots represent model-based predicted response values at eight time points, at each of which a Bayesian ANOVA is applied (values at three time points are circulated). The experimental group represented by black colour is considered as a control group, against which the others are compared to. Thus, temperature effect refers to the difference between the black and green curves, and strain effect to the difference between the black and red curves. If no interaction between temperature and strain exists, equal effect as is present between temperatures for strain 1 (black and green) would also be present between temperatures for strain 2 (red and blue). Similarly, equal effect as is present between strains at low temperature (black and red) would also be present between strains at high temperature (green and blue). Right: Estimated effects with their 95% credible intervals. An effect is considered statistically significant if its credible interval does not contain zero. We see that temperature effect is significantly positive throughout the experiment (green), whereas strain effect (red) is significant only at the beginning and at the end, and changes its direction from positive to negative in between. Since the interaction (blue) is significant, the temperature effect present in strain 1 is not present in strain 2, and similarly the strain effect at high temperature is different from the (mainly non-existing) effect at low temperature.
2.1.2 Methods introduced in article II

Article II introduces another pipeline called FLOG for studying Biolog metabolic curves. The approach is designed for visual inspection rather than for statistical testing of metabolic differences. However, the effects detected by FLOG can be tested, for example, by using the Bayesian ANOVA approach introduced in article I.

The method is motivated by the multidimensional nature of the Biolog data, and designed for comparing several samples. In comparison to the traditional methods described above and, for example, in [38], this method does not require replicates. The outline of the approach is similar to the
framework demonstrated in Figure 2.3, with the exception that three summary statistics are used instead of a single value. The statistical basis of the method is factor analysis rather than PCA or MDS.

The aim was to implement a straightforward pipeline for getting an overview of vast amounts of Biolog data by reflecting the originally multidimensional data into a two- or three-dimensional factor space, and clustering the samples based on their factor loadings. Additionally, each single curve or the mean metabolic curve of the samples within a cluster is plotted for a visual inspection of either similarities or differences within or between the clusters. Based on our experience, the tool nicely enables the detection of outlying samples, replicates, or other groups with dissimilar metabolic patterns. The advantage of using FLOG is that features specific to Biolog data, such as background correction, normalization, and exclusion of inactive wells, are taken into account in the pipeline. Also several PM plates can be simultaneously analyzed. Furthermore, FLOG is compatible with the pipeline introduced in article I, thus in addition to performing factor analysis, the data can be grouped, normalized and compared with the existing R functions.

\[
\text{FLOG} = \text{Factor analysis} + \text{BioLOG}
\]

**Multidimensional Biolog data** = Biolog data are multidimensional by nature as thousands of phenotypes are tested simultaneously, and additionally at hundreds of time points

**Factor space** = a space defined by uncorrelated factors that depicts the most of the variation in the observed data points

**Factor loading** = the strength of the relationship between a data point and a factor
Figure 2.6: Reflection of the originally multidimensional Biolog data into a three-dimensional factor space. Factor analysis is applied on the summary statistics of the metabolic activity curves represented in Article II. Data include 98 samples, of which 18 are categorized as wild-type samples and 80 as single gene knock-out mutants. The resulting factor solution is represented as three-dimensional plots, each of which show the same factor solution with varying clustering. Top: factor loadings clustered by k-means with $k = 3$. Middle: factor loadings clustered by k-means with $k = 4$. Bottom: factor loadings coloured according to the wild-type/mutant status of the samples. The fading colours and the tails visualize the position of the loadings on the third factor.
Chapter 3

Genome-wide association studies (GWAS)

The genetic basis of each living cell is the DNA (deoxyribonucleic acid) molecule, a sequence of nucleotides alternating between four bases: cytosine (C), guanine (G), adenine (A), and thymine (T). The size of an organism’s genome is generally considered as the total number of bases. For a human, the genome size is about 3 billion ($3 \times 10^9$) bases, and the order of the nucleotides is highly constrained. In fact, the DNA of any two persons is $99.5\% - 99.9\%$ identical, depending on the source [42, 43]. Yet, the subtle variation that remains, makes each person unique with different hair colours, facial structures, and other traits. The small differences in our genomes are also a key factor in tracking causes of some diseases, such as hereditary breast, and colorectal cancer, or Parkinson’s and Crohn’s disease [44, 46–49]. Genome-wide association studies (GWAS) provide a tool for identifying such genetic variants involved in the development of human diseases and individual traits.

The most common type of genetic variation among people is single nucleotide polymorphism (SNP), found on average once in every 300 nucleotides [50]. As the name suggests, SNP is a single nucleotide difference at a certain base position of the DNA sequence. For example, at a SNP, the nucleotide C may be replaced with the nucleotide T. Most SNPs so far discovered include only two variants (bi-allelic SNP), but there are also SNPs in which three different base variations coexist [51]. In bi-allelic SNPs one variant is present less frequently than the other, and the frequency of the less frequent allele is referred as the minor allele frequency (MAF). The human genome contains roughly 10 million SNPs for which the MAF is 5\% or more, i.e. both variants are found in at least 5\% of individuals in the human population.
SNPs are not the only type of genetic variation, as human genome also contains larger regions that are deleted, duplicated, repeated, or inverted. Such genetic markers are called copy-number variations (CNVs). However, when applying GWAS, SNPs are preferred over other genetic markers, because of their high abundance, relatively low mutation rate, and easy adaptability to automatic genotyping.

At its simplest, GWAS tests for non-random distribution of nucleotides between two groups of people: cases and controls. The testing is most often done one SNP at a time, and a SNP is considered as causal for a particular disease if it occurs significantly more frequently in people with the disease (case) than in people without the disease (control), and it introduces, for example, a nonsynonymous change to a coding sequence or alters gene regulation. Ideally, the controls are matched as closely as possible with the cases in other respects than in terms of the disease status. Table 3.1 shows an example distribution for two alleles at a single SNP. Whether either of the alleles occurs more frequently among the cases than controls, can be tested, for instance, by \( \chi^2 \)-test.

<table>
<thead>
<tr>
<th>Allele</th>
<th>Disease status</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>case</td>
<td>control</td>
<td>total</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>23</td>
<td>11</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>73</td>
<td>15</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>96</td>
<td>26</td>
<td>122</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Distribution of alleles at a bi-allelic SNP.

As the human genome contains millions of SNPs, it would be difficult, time-consuming, expensive and ineffective (in terms of statistical power) to look for changes in each of these. Therefore, genotyping effort is greatly reduced, as well as the statistical power increased, by utilizing the assumption that SNPs close to each other in a chromosome are inherited together, forming correlated clusters of SNPs known as haplotype blocks in which SNPs are predictive of each other [52]. Thus, instead of measuring millions of SNPs, it is sufficient to collect a smaller fraction of tag SNPs from the haplotype blocks. Also it is easier to detect loci associated with a phenotype, especially if there are many SNPs strongly correlated with the causal SNP. The phenomenon of correlated loci is also known as linkage disequilibrium (LD).

Once a disease-causing SNP is detected, strategies for diagnosis or pre-
vention of the disease can be suggested. For example, the discovery of a SNP associated with lactase persistence provided a new approach for diagnosing patients with lactose intolerance: the earlier used, uncomfortable examination provoking symptoms of lactose intolerance could be replaced with a genetic test which can easily be taken from a blood sample [4,17]. Another example is the prevention of hereditary breast cancer caused by pathogenic variants in \textit{BRCA1} or \textit{BRCA2} genes [44]. The options for prevention are regular breast screenings, a surgery to remove breasts (and possibly ovaries), or medicines to lower the risk of developing cancer.

In addition to causing diseases, or increasing the risk of them, differences in genetic composition can affect individuals’ response to pathogens, chemicals, drugs, vaccines, and other agents, thus allowing the design of tailored treatment for patients according to individual genetic features [45].

| **Nucleotide** = a biological compound of a nitrogenous base (A, C, G, T), sugar, and a phosphate group |
| **Nonsynonymous change** = a nucleotide mutation in the DNA sequence that alters the resulting amino acid sequence of a protein in contrast to not altering which would then be considered as a synonymous change |
| **χ²-test** = a statistical hypothesis test to determine whether there is a significant difference between the expected and the observed frequencies in one or more categories of a contingency table |
| **Statistical power** = the ability of a test to detect an effect, if the effect actually exists, \textit{i.e.} low power means that real effects remain undetected |

**Concerns related to GWAS**

There are some complications related to performing genome-wide association studies. Starting from scratch, both \textit{genome sequencing} and \textit{SNP calling} are delicate multi-step processes and as such prone to errors. Thus, the accuracy of the techniques used in sequencing and defining the single nucleotide differences between sequences, directly reflects to the downstream genomic analyses, such as GWAS [53,54].

For many common disorders, like heart diseases, diabetes or cancer, there exists no single SNP sufficient to cause the disease, \textit{i.e.} these diseases are due to the combined effect of several genetic variants, making the genome-wide analysis more complicated [55]. Additionally, other factors, such as environment, have a significant role in the development of many diseases, such as lung cancer [56].

Some diseases are rare, and thus difficult to study in terms of collecting
large enough sample sizes. Some genetic variants are rare as well. When only a few people in a study carry a variant, large sample sizes are required to detect patterns in the genetic composition. Some rare variants are quite young, and partly due to the increased opportunities for mutations to occur during the rapid growth of human population since the beginning of the 20th century [57,58]. In some diseases, both rare and common genetic variants have a combined effect on the risk of the disease, thus complicating the analyses [59].

The advantages of haplotype blocks and LD were discussed above, however LD also introduces some problems to the genome-wide analysis of genetic variants, as it might be difficult to locate the causal SNP among the correlated SNPs, especially if there are many SNPs in strong LD with the causal SNP [60].

Yet another problem is introduced by sub-populations, as a SNP allele common in one geographical or ethnic group may be much rarer in another. Therefore, causal SNPs found by GWAS for any phenotype that varies across sub-populations are likely predictive of the individual’s sub-population of origin rather than the studied phenotype. Clearly such population structures need to be accounted for in the analysis to avoid false positive discoveries. However, the underlying population structures are not always known in advance, and thus need to be estimated.

Finally, the testing of thousands of SNPs at a time introduces a problem of multiple hypothesis testing, reducing the power of detecting the truly causal SNPs.

<table>
<thead>
<tr>
<th>Genome sequencing</th>
<th>= reading the bases of a DNA sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP calling</td>
<td>= finding sites that vary between the compared genome sequences</td>
</tr>
<tr>
<td>Rare (genetic variant)</td>
<td>= a genetic variant, for example, a SNP that occurs at a low frequency in a population</td>
</tr>
<tr>
<td>Multiple hypothesis testing</td>
<td>= considering a set of statistical hypothesis simultaneously making it more likely to find statistical significance by random chance alone</td>
</tr>
</tbody>
</table>

### 3.1 GWAS in bacteria

In terms of bacteria, researchers are interested in finding genetic variants associated with a wide variety of phenotypes, such as host, geographical area, pathogenicity or lineage. Currently, especially intriguing is solving the genetic factors behind antibiotic resistance of bacteria, as many human pathogens have developed the ability to adapt and overcome antibiotics,
complicating the treatment of diseases caused by bacteria [61]. Many bacteria are resistant to multiple types of antibiotics, even to all available antibiotics, as well as can develop resistance during an ongoing treatment. Examples of such multi-drug resistant species are methicillin-resistant *Staphylococcus aureus* (MRSA) which is a major source of hospital-acquired infections, *Pseudomonas aeruginosa* and *Mycobacterium tuberculosis*.

However, association studies are not as widely applied for bacteria as they are for human [18–20]. This is due to various reasons, for instance, it has become sufficiently affordable to sequence hundreds or thousands of isolates from a bacterial population only very recently, and before this population genomic era association studies in bacteria were restricted to candidate genes. Other complications arising in bacterial GWAS are the clonal population structure and restricted recombination in bacterial populations [18,19,62]. The clonal population structure is due to an asexual reproduction process in which an offspring inherits the genes of its parent, resulting in offsprings identical, or nearly so, to their parents. Human gamete cells, instead, are produced by meiosis in which recombination, *i.e.* exchange of genetic material between chromosomes occurs naturally, resulting in offsprings whose DNA content differs from that of the parent. Since the effectiveness of GWAS in detecting genetic relationships depends crucially on the degree of variation introduced in the compared sequences by recombination, association studies are of limited value for completely clonal or infrequently recombining organisms. In other words, in bacteria recombinations occur too rarely to break the genetic material into blocks where the causal parts could be distinguished from the non-causal ones by computational means. Additionally, strong population structure in bacteria can produce false positive discoveries and loss of statistical power in detecting causal variants.

Even though bacterial homologous recombination is restricted in general, it does occur via several molecular mechanisms. This process is often called lateral or horizontal gene transfer (LGT/HGT), but it may also more generally introduce non-homologous DNA and alter the gene content of a bacterial chromosome. The main modes of HGT are typically categorized into three different processes in which a living cell can uptake DNA from its surroundings (transformation), from other bacteria (conjugation), or with the assistance of bacteriophages (transduction) [63,64]. Although the study of recombination is restricted due to the clonal structure of bacterial populations [65,66], the recombination rates have been detected to vary considerably across species [64,67]. *Streptococcus pneumoniae, Pseudomonas aeruginosa* and *Helicobacter pylori* are examples of species with
a high recombination rate, *Burkholderia pseudomallei* and *Campylobacter jejuni* have intermediate recombination rate, and *E. coli* is typically associated with a low homologous recombination rate. It should be remarked that despite of generally low levels of homologous recombination, an *E. coli* population still may be experiencing gene gain and loss through HGT at a very high rate as exemplified by the recent study of the global pandemic clone ST131 [21].

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Host</strong></td>
<td>source of bacteria, e.g. human, cattle, poultry</td>
</tr>
<tr>
<td><strong>Human pathogen</strong></td>
<td>a virus, bacterium, prion, or fungus that causes disease in humans</td>
</tr>
<tr>
<td><strong>Asexual reproduction</strong></td>
<td>the primary form of reproduction for single-celled organisms which does not require fusion of gametes</td>
</tr>
<tr>
<td><strong>Gamete</strong></td>
<td>a mature sexual reproductive cell, such as sperm or egg, that unites with another cell to form a new organism</td>
</tr>
<tr>
<td><strong>Population structure in bacteria</strong></td>
<td>diverse genetic composition within or between bacterial species due to recombination or mutations</td>
</tr>
<tr>
<td><strong>Lateral/horizontal gene transfer (LGT/HGT)</strong></td>
<td>a mode of recombination in bacteria to transmit DNA between different genomes, for example, in order to spread a beneficial gene that produces more durable organisms, or to ensure the survival of a species</td>
</tr>
<tr>
<td><strong>Homologous recombination</strong></td>
<td>a major DNA repair process in bacteria in which nucleotide sequences are exchanged between two similar or identical molecules of DNA</td>
</tr>
<tr>
<td><strong>Non-homologous recombination</strong></td>
<td>a type of recombination in which a nucleotide sequence is translocated into a new position in a genome</td>
</tr>
</tbody>
</table>

### 3.1.1 Sequence element enrichment analysis (SEER)

In article III, we introduce a new approach, sequence element enrichment analysis (SEER), for applying bacterial genome-wide association analysis. Instead of SNPs, we test whether sequence elements, more generally known as k-mers, are over- or under-represented within a phenotype. K-mers, *i.e.* DNA words of length k, have been used for numerous purposes in bioinformatics, for instance for assembling sequencing reads to contigs [68], and to building alignment-free phylogenies [69]. The benefit of k-mers over SNPs in GWAS is that they capture several different types of variation present in genomes as well as expand the search of genetic variants to accessory genomic regions instead of only focusing on the core genome.

SEER is compiled into a pipeline which provides several alternatives for the input data, counting k-mers, performing analysis, and representing the results. The possibly underlying population structures are revealed
by estimating the genetic distances between the isolates by constructing
a distance matrix based on a random subset of k-mer occurrences. An
example of the binary k-mer occurrence vectors where the presence of a
single k-mer is recorded as a binary variable, i.e. present or absent, is shown
in Table 3.2. The table also illustrates, how the pairwise distances are
calculated. Multidimensional scaling (MDS) is then applied on the distance
matrix constructed of the pairwise distances, and the eigenvectors of the
MDS projection used as covariates when testing for associations. If three
eigenvectors explain the variation in the data, the population structure can
be represented as a three-dimensional plot as illustrated in Figure 3.1.

Table 3.2: K-mer occurrence vectors of length \( m \) for \( n \) isolates
represented in a matrix.

<table>
<thead>
<tr>
<th>Isolate</th>
<th>k-mer ( 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ldots \ m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 0 0 1 0 0 1 \ldots 1</td>
</tr>
<tr>
<td>2</td>
<td>1 1 0 0 0 1 0 1 \ldots 1</td>
</tr>
<tr>
<td>3</td>
<td>0 1 0 0 1 0 0 1 \ldots 0</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>n</td>
<td>0 0 0 0 0 0 1 1 \ldots 1</td>
</tr>
</tbody>
</table>

Each row in the matrix represents an occurrence vector for a single isolate. A random subset of
0.1 - 1 \% of all the k-mers is used, and for each presence in a particular sample coded by 1, and
absence by 0. Pairwise Hamming distances are calculated and collected into a matrix (matrix
representation is not shown here).

The phenotype, such as response to antibiotic treatment, is also con-
sidered as binary. In a similar fashion to SNPs (Table 3.1), the data for
k-mers can be represented as a \( 2 \times 2 \) contingency table as shown in Table
3.3.

The association of the antibiotic resistance to the presence or absence
of a k-mer can be tested by applying a logistic regression model:

\[
\log \left( \frac{y}{1-y} \right) = \beta_0 + \beta_1 x + \beta_2 Z, \tag{3.1}
\]

where \( y \) is a binary outcome vector coding for the antibiotic response (1
if resistant, 0 if sensitive), \( x \) a binary vector coding for the k-mer presence
(1 if present, 0 if absent), \( Z \) a matrix containing the eigenvectors of the
Table 3.3: **Distribution of k-mers across resistant and sensitive samples.**

<table>
<thead>
<tr>
<th>K-mer presence</th>
<th>Antibiotic response</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>resistant</td>
<td>sensitive</td>
<td>total</td>
<td></td>
</tr>
<tr>
<td>absent</td>
<td>23</td>
<td>11</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>present</td>
<td>73</td>
<td>15</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>96</td>
<td>26</td>
<td>122</td>
<td></td>
</tr>
</tbody>
</table>

MDS projection and additional user-given covariates, and β’s include the regression coefficients (β₀ and β₁ are scalars and β₂ is a vector). The effect of the k-mer presence on the antibiotic response is evaluated by testing whether β₁ = 0. If β₁ < 0, the k-mer is considered under-represented within the resistant samples, and similarly if β₁ > 0, over-represented.

In logistic regression, rare outcome events, *i.e.* if in either of the phenotype groups, the tested k-mer is found in few or none of the samples, or it is found in nearly all of the samples, cause problems in fitting the model which can be detected as large standard errors. A similar event occurs when the continuous population structure covariates predict the outcome phenotype too perfectly, or in other words, the population structure is strongly correlated with the phenotype (see Figure 3.1). This phenomenon is known as (perfect) separation, and is overcome in SEER by invoking Firth regression [70] in which the likelihood function is penalized by a factor known as Jeffreys prior, the effect of which diminishes as sample size increases:

\[ L(\beta)^{\text{Firth}} = L(\beta) |I(\beta)|^{\frac{1}{2}}, \tag{3.2} \]

where \( L(\beta) \) is the likelihood function in terms regression parameters \( \beta \) and \( |I(\beta)|^{\frac{1}{2}} \) the penalizing factor, where \( |I(\beta)| \) is the determinant of the information matrix evaluated at \( \beta \).

In general, logistic regression can be utilized when the outcome variable, *i.e.* phenotype, is dichotomous, and the predictor variable, *i.e.* genotype, is categorical, or continuous. So far, only dichotomous phenotypes have been discussed, however SEER also incorporates the analysis of continuous phenotypes by applying linear regression instead of the logistic regression [3.1].

A typical dataset analyzed by SEER contains hundreds of samples, each of which is a DNA sequence millions of bases long, thus yielding millions of k-mers to be analyzed. Typical to GWAS, all the variant sites, whether
SNPs or k-mers, are tested one at a time, creating the problem of multiple hypothesis testing as well as an enormous computational burden. SEER tackles these problems by setting a strict cut-off level for p-values, using an effective C++ coding environment, reducing the number of k-mers tested with the logistic model [3.1] by pre-filtering the k-mers with a $\chi^2$-test, and using a step-wise approach when solving the model parameters, i.e. a faster, but more error prone algorithm is first applied, and only if needed a more successful, but more time consuming option. As discussed above, problems in fitting the logistic model caused by separation are overcome by adding an adjustment to the likelihood function when solving the regression coefficients (see Equation 3.2).
3.1.2 Applications of SEER to real data

We have applied SEER to over a dozen of bacterial studies with a variety of phenotypes of interest, such as host, age of the host, geographical location, antibiotic resistance, and invasiveness, and have been able to discriminate genetic variants specific to these features. Table 3.4 summarizes some of these applications by specifying the bacterial species studied accompanied with the length of its genome, the tested phenotype, the number of samples per group and in total, as well as the number of distinct and significant k-mers of length between 9 and 100.

As indicated in Table 3.4, the most studied genus is Streptococcus for which we have investigated the species: Streptococcus suis, Streptococcus pneumoniae and Streptococcus pyogenes. The length of a bacterial genome is, in general, some millions of base pairs varying here roughly from two to six million base pairs. In comparison to the human genome, the bacterial genome is about 1000 times smaller. Among the species studied, Streptococci and Campylobacter jejuni have the shortest genomes of about two
Table 3.4: Several of the SEER applications considered by the author.

<table>
<thead>
<tr>
<th>Species</th>
<th>Ref.</th>
<th>Genome size</th>
<th>Phenotype</th>
<th>Sample size</th>
<th>K-mers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>size (^{a})</td>
<td></td>
<td>Group</td>
<td>Total</td>
</tr>
<tr>
<td><em>Streptococcus suis</em></td>
<td>Article IV</td>
<td>2.1</td>
<td>Host</td>
<td>38/153</td>
<td>191</td>
</tr>
<tr>
<td><em>Streptococcus pneumoniae</em></td>
<td>Article III</td>
<td>2.3</td>
<td>Antibiotic resistance</td>
<td>179/2890</td>
<td>3069</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>585/2484</td>
<td>3069</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1171/1898</td>
<td>3069</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>768/2301</td>
<td>3069</td>
</tr>
<tr>
<td><em>Streptococcus pyogenes</em></td>
<td>Article III</td>
<td>1.9</td>
<td>Invasiveness</td>
<td>185/489</td>
<td>674</td>
</tr>
<tr>
<td><em>Pseudomonas aeruginosa</em></td>
<td></td>
<td>6.3</td>
<td>Antibiotic resistance</td>
<td>40/177</td>
<td>217</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>29/188</td>
<td>217</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>42/175</td>
<td>217</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>42/175</td>
<td>217</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>88/129</td>
<td>217</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>87/130</td>
<td>217</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11/202</td>
<td>213</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>89/127</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>79/137</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100/116</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>72/144</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>80/137</td>
<td>217</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12/201</td>
<td>213</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40/177</td>
<td>217</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>38/179</td>
<td>217</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21/190</td>
<td>211</td>
</tr>
<tr>
<td><em>Burkholderia pseudomallei</em></td>
<td>[22]</td>
<td>3.2</td>
<td>Geographical location</td>
<td>11/290</td>
<td>301</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9/292</td>
<td>301</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9/292</td>
<td>301</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>63/238</td>
<td>301</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>103/198</td>
<td>301</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>[21]</td>
<td>5.0</td>
<td>Multi-drug resistance</td>
<td>227/722</td>
<td>949</td>
</tr>
<tr>
<td><em>Campylobacter jejuni</em></td>
<td>[23]</td>
<td>1.7</td>
<td>Geographical location</td>
<td>141/175</td>
<td>316</td>
</tr>
<tr>
<td><em>Salmonella</em></td>
<td></td>
<td>4.7</td>
<td>Lineage</td>
<td>245/252</td>
<td>497</td>
</tr>
<tr>
<td><em>H1N2</em></td>
<td></td>
<td>0.014</td>
<td>Vaccination</td>
<td>112/183</td>
<td>318</td>
</tr>
</tbody>
</table>

The table summarizes some of the applied SEER applications by specifying the species studied, a reference to the SEER study, an approximate length of the genome studied, the phenotype tested, the number of samples per group and in total, as well as the number of distinct and significant k-mers found. The categorization for the significant k-mers is the following:

Number of significant k-mers found

- 0
- 1 – 10\(^{2}\) \(\text{I}\)
- 10\(^{2}\) – 10\(^{3}\) \(\text{II}\)
- 10\(^{3}\) – 10\(^{4}\) \(\text{III}\)
- 10\(^{4}\) – 10\(^{5}\) \(\text{III}\)
- > 10\(^{5}\) \(\text{IIII}\)

\(^{a}\) Genome size in millions base pairs; Source: http://www.kegg.jp/kegg/genome.html

\(^{b}\) Prefiltered

\(^{c}\) Virus

The table summarizes some of the applied SEER applications by specifying the species studied, a reference to the SEER study, an approximate length of the genome studied, the phenotype tested, the number of samples per group and in total, as well as the number of distinct and significant k-mers found. The categorization for the significant k-mers is the following:

The most effective study design million base pairs, while *Pseudomonas aeruginosa*, *E. coli* and *Salmonella* have the longest genomes of about five to six million base pairs.

All the phenotypes tested are dichotomous, meaning that the logistic regression model [3.1] was applied in SEER when searching for k-mers enriched in either of the phenotype groups. The sample sizes vary from 191 to 3069, being on average some hundreds. The most effective study design
would contain an adequate and equal number of samples in both of the phenotype groups. However, as can be observed, the settings rarely are balanced, mainly due to the phenotype of interest being rare. For example, not many samples were available from the same geographical location in the *Burkholderia pseudomallei* study. A mild unbalance may not affect the results, but if either of the groups has too few observations, it decreases the power of detecting enrichment as well as increases the number of false positive signals.

The number of distinct k-mers found is dependent on, *e.g.*, the length of the genome, number of the samples, length of the actual sequences provided, spectrum of k-mer lengths (in these data sets k-mers between 9 and 100 were extracted) and the overall similarity between the compared strains. For some studies, k-mers were filtered during scanning, hence decreasing the number of k-mers found. In Table 3.4, the number of distinct k-mers varies between 550,000 and 100 million (19,000 for *H3N2* virus).

The number of significant k-mers represented in Table 3.4 is the number of k-mers before applying any additional filtering which may be needed if the number of significant k-mers is vast, like for *Streptococcus pneumoniae* and *E. coli*. Filtering reduces the amount of work when mapping the k-mers back to a reference, and tracking the functions as well as the locations of the k-mers. For most of the bacteria studied, SEER identified k-mers enriched within the phenotypes tested, *i.e.* genetic variants associated with antibiotic resistance, pathogenicity, geographical areas, multi-drug resistance and different lineages seem to exist. For *Streptococcus pneumoniae*, the antibiotic resistance determinants identified by SEER are previously characterized [18], thus confirming the performance of SEER in detecting sequence elements associated with phenotypes.

For some bacteria, the k-mer contents do not vary between the phenotype groups, *i.e.* no enrichment or in other words genotype-phenotype associations are detected. For example, in the *Streptococcus suis* study, represented in article IV, the genetic composition of isolates extracted from human samples were compared against isolates from pigs. *Streptococcus suis* is known as a swine pathogen, and of interest was to address the adaptation of the bacteria to the human population in Vietnam. The complete data include isolates also from United Kingdom and China, but in this part of the study, only the 191 isolates collected from Vietnam (38 from pigs and 153 from humans) were used. An initial phylogenetic analysis performed based on the complete data set and a discriminant analysis of principal components based on the 191 Vietnamese isolates represented in Figures 3.2 and 3.3, respectively, show no structuring of the isolates by
3.1 GWAS in bacteria

host, indicating no consistent adaptation to the human population. To verify this hypothesis genome-wide association analyses were applied on the 191 Vietnamese isolates, both based on SNPs and k-mers, and, indeed, no significant associations were detected, i.e. no genetic variants, neither SNPs nor sequence elements, appeared significantly more often in isolates from human than pig infections. It should be noted, however, that the study had relatively low power due to a small number of genomes so it cannot be excluded that host adaptive elements could be detected in larger follow-up studies in the future.

Pathogenic abilities of the studied species

Most of the studied bacteria unlikely cause severe infections to healthy people, however people with other illnesses or conditions, such as patients receiving chemotherapy for cancer, undergoing a surgery, having cystic fibrosis, HIV or AIDS, mechanical ventilator or catheter, or burn wounds, are at a higher risk due to weakened immune response. Also, young children and elderly people are at increased risk of bacterial infections.

*Streptococci*, in general, harbour the respiratory tract and skin of healthy hosts without signs of disease, but when introduced to vulnerable tissues can cause severe infections. *Streptococcus suis* is known as a swine pathogen, causing respiratory tract infections in pigs as well as invasive diseases, such as arthritis, septicemia and meningitis [71], but it can also infect humans. *Streptococcus pneumoniae* or the *pneumococcus*, is instead a human pathogen causing both mild (bronchitis, ear infection, sinusitis) and severe (blood poisoning, pneumonia, meningitis) infections. *Streptococcus pyogenes* has an ability to cause suppurative infections, mainly on skin, but it also causes infections in throat, such as angina.

*Pseudomonas aeruginosa* is found in moist areas, such as soil and water, and can be the cause of skin, ear, or eye infection originated, for instance, from an inadequately chlorinated swimming pool. The more severe infections induced by *P. aeruginosa* are bacteremia (infection in blood) and pneumonia (infection in lungs). *P. aeruginosa* is also a fairly common pathogen involved in infections acquired in hospital settings.

*Burkholderia pseudomallei*, also found in soil and water, causes Melioidosis with pain in chest, bones, or joints, cough, skin infections, lung nodules and pneumonia [22]. Melioidosis is endemic, particularly to Thailand and northern Australia.

*Escherichia coli* naturally colonizes the lower intestine of humans and animals, and is probably the most studied, and hence the best-known bacterial species. Most *E. coli* strains are harmless, but some are pathogenic causing serious food poisoning due to contamination of food or water by fecal matter as well as urinary tract infections, bloodstream infections, and meningitis. In developing countries, *E. coli* is a common cause of infant diarrhea. Recently, sequence type ST131 has emerged as a major cause of serious multidrug-resistant extraintestinal *E. coli* infections [21].

*Campylobacter jejuni* and *Salmonella* are, in addition to *E. coli*, common causes of food-borne intestinal infections [23].

*H3N2* is a seasonal influenza virus.
Figure 3.2: Phylogenetic tree of a virulent zoonotic clade of *Streptococcus suis*. The tree shows 256 human and pig isolates collected from Vietnam, United Kingdom and China, and is estimated from an expanded core genome. Terminal branches are coloured according to the country and host of origin, and indicate some genetic structuring by country, but little clustering by host. Especially scattered are the isolates extracted from pigs from Vietnam. The clade denoted with an asterisk corresponds to the isolates shown in the right hand peak of Fig 3.3. Figure from article IV.
Figure 3.3: Discriminant analysis of principal components, applied to 191 isolates collected from human and pig hosts in Vietnam. The analysis is applied to SNPs in the core genome, and to presence/absence data for genes in the accessory genome. The first linear discriminant function is shown, and the lack of separation between the distributions suggests a lack of consistent genetic differences between isolates from the two host types. Figure from article IV.
3 Genome-wide association studies (GWAS)
Chapter 4

Discussion

In this thesis, the aim was to develop statistical analysis tools for two types of data produced by modern high-throughput technologies. On one hand Biolog phenotype microarray technology along with the most commonly used, as well as novel, analysis methods were introduced; on the other hand a new genome-wide association analysis approach, SEER, was proposed to bacterial sequence data. Both settings and technologies address their own requirements which need to be taken into account to be able to produce adequate results. Additionally, easy to use pipelines and software are fundamental in handling and processing the vast amounts of data.

As both methodologies produce a lot of information, dimension reduction methods are applicable and useful. These methods have originally been developed in social sciences and psychometry [73,74], but have more recently started to appear in molecular biology, for instance, for grouping genes in DNA microarray analysis [75,76]. In the context of genome-wide association analysis, we utilize dimension reduction for estimating the population structure, whereas in the context of Biolog phenotype microarrays, these methods help to visualize the data and reduce the need for multiple hypothesis testing, the problem of which is also present in genome-wide association studies. SEER tries to diminish the effect of testing multiple hypothesis by limiting the number of tests by pre-filtering as well as by applying a strict cut-off for p-value.

Similar to any statistical testing, the data types covered in this thesis require consideration of variables not included or controlled in the study. In general, such characteristics are called confounding variables, and they are known to affect the results by clouding the real effects [72]. In the context of phenotype microarrays, laboratory conditions, the technician used, or the number of cells pipetted are examples of such confounding factors, whereas in the genome-wide association studies, the most crucial confounding factor

35
is the population structure.

Well designed experiments control for confounding variables in advance, for instance, by using the same equipment and laboratory personnel, matching pairs of subjects based on the potential confounding factors and randomly assigning the subjects to treatment groups. However, such control is not always possible, for example, laboratory conditions are not identical if metadata are collected from several sources, or subjects can not be randomly assigned to have a disease status, for instance, when studying whether the genomes of bacteria isolated from cancer patients differ from non-cancer patients. If the effects of confounding factors are not diminished when designing the experiment, they can still be taken into account in the statistical analyses by measuring the variables and including them as covariates in a statistical model. However, these variables are not always, or cannot be, measured. Therefore they often need to be estimated, as is done in terms of the population structure in genome-wide association analyses. In SEER, the population structure is estimated as well, and included as a covariate in the regression model. In the context of Biolog phenotype microarray experiments, confounding factors rarely are recorded either. Therefore, our pipeline looks for hints of confounding effects by comparing replicates. Then the effects are diminished by applying a procedure called normalization.

The future prospects in terms of the Biolog and genomic data include combining the analysis of the two types of data. One approach to link metabolic data with genome sequence data was suggested by Galardini et al. [78]. It separately analyzes both the genetic and metabolic data, and represents the results on the top of KEGG pathways. Another valid approach would be to first look for genetic variants associated with a certain disease or state by genome-wide association studies, then imply the revealed mutation(s) into a cell, and address their metabolic consequences with Biolog phenotype microarrays. However, the most intriguing approach would consider Biolog data as the phenotype data in the genome-wide association study. This would allow to detect correlations between the metabolic activity and the genetic composition of an organism in a wider scale. At the moment, such approaches are limited by the small amount of Biolog data available as well as the multidimensional nature of the Biolog phenotype data. Ideally, the association of multiple phenotypes with genetic factors should be addressed at once, and not one phenotype at a time. Some high-dimensional phenotypes have already been utilized in genome-wide association studies [77], thus as soon as the sample sizes increase GWAS should be applicable to Biolog phenotype microarray data as well.
References


