Universität des Saarlandes
Zentrum für Bioinformatik

Data Mining Techniques for improving and enriching cancer epigenetics

Dissertation im Fach Bioinformatik
PhD Thesis in Bioinformatics
von / by

Dr.-Ing. Ruslan Akulenko

angefertigt unter der Leitung von / supervised by

Prof. Dr. Volkhard Helms

begutachtet von / reviewers

Prof. Dr. Volkhard Helms

Prof. Dr. Tobias Marschall

Saarbrücken, April 2016

Day of Colloquium	01 / 04 / 16
Dean of the Faculty	Univ.- Prof. Dr. Frank-Olaf Schreyer
Chair of the Committee	Prof. Dr. Gert Smolka
Reporters	
First reviewer	Prof. Dr. Volkhard Helms
Second reviewer	Prof. Dr. Tobias Marschall
Academic Assistant	Dr. Daria Stepanova

Acknowledgements

This current work was accomplished in the Center for Bioinformatics at the University of Saarland. Firstly, I would like to thank my supervisor Prof. Volkhard Helms for his helpful advices and giving me the opportunity to carry out my PhD studies in hist group. Furthermore, I would like to thank our collaborators Prof. Mathias Herrmann, Prof. Lutz von Muller, Ulla Rufing and other members of African-German Network on Staphylococci for very productive joint work and introducing to me the analysis of bacterial genome data. A special word of thanks goes to Pavlo Lutsik with whom we have successfully advised master student in our group.
Finally, I would like to thank all members of Prof. Helms group and my wife Tetiana for supporting me during my studies.

Abstract

The aim of epigenetic cancer research is to connect tumor development with processes involving gene regulations, which are not directly encoded in the DNA. Recent advanced techniques for genome-wide mapping of epigenetic information are significantly improving this knowledge by generating large amounts of high resolution data. But still many dependencies remain unknown due to the presence of technical errors such as batch effect. This thesis describes the associations discovered during DNA methylation data mining as well as technical challenges which should be considered during data processing. Moreover, the same computational methods applied to epigenetic data were adapted to process bacterial genome data.

Hereby, thesis is mainly structured into four parts. The first part presents the baseline for DNA methylation data analysis, pathway and functional enrichment which helped in explaining DNA co-methylation phenomenon. The second part covers the problem of batch effect detection and adjustment. Here the newly developed BEclear R software package is described and is compared against existing well established methods. In addition, we suggest the optimal strategy for batch effect assessment, parameter selection and the impact of this negative effect on the final result. The third part addresses the analysis of *S. aureus* genome using different data mining techniques ranging from hierarchical agglomerative clustering to Affinity Propagation. Interestingly, that ambiguous bacterial genome data was successfully predicted with the same exact matrix completion algorithm used in BEclear. The final part describes statistical analysis of community acquired *S. aureus* isolates which revealed important associations between bacterial resistance and virulence profiles with the geographical location.

Conceivably the most important outcome of the current thesis is the in-depth review of batch effect. It is notorious for its ability to affect the whole data processing procedure. Moreover, it can influence the results tremendously and is of current interest among researchers who work with genome-wide high-throughput data sets. There exist several approaches that allow diminishing the negative effect of batch effect on the investigated data sets. Our new tool BEclear allows not only detecting and assessing batch effect, but also adjusting it only in the batch affected part of data using Latent factor models matrix approximation. We tested the devised methodology on breast invasive carcinoma data from The Cancer Genome Atlas and compared it with the existing algorithms ComBat, Surrogate Variable Analysis and Functional normalization. We show that BEclear outperformed these methods with respect to precision while avoiding changing the unaffected data, since it focuses on the batch affected genes only. This makes BEclear a competitive algorithm in batch effect correction and can be widely applied to DNA methylation or even gene expression data. BEclear is available as an R package.

Kurzfassung

Das Ziel der epigenetischen Krebsforschung liegt darin, die Verknpfungen der Tumorentwicklung mit der Genregulation, die nicht direkt durch die DNA kodiert ist, aufzudecken. Vor kurzem erfolgte Fortschritte in den Techniken fr die genomweite Kartierung epigenetischer Informationen steigerten den Wissensstand erheblich durch die Erzeugung riesiger, hochaufgelter Datenmengen. Dennoch bleiben viele Abhngigkeiten, aufgrund des Vorhandenseins von technischen Fehlern wie etwa von Batch Effekten, unentdeckt. Diese Doktorarbeit beschreibt die Zusammenhnge, die whrend der Gewinnung von DNA-Methylierungsdaten entdeckt wurden, sowie technische Herausforderungen, die bei der Datenverarbeitung bercksichtigt werden sollten. Darber hinaus wurden die gleichen computergesttzten Methoden, die zur Verarbeitung der epigenetischen Daten angewandt wurden, zur Verarbeitung von Bakteriengenomdaten angepasst.
Diese Abschlussarbeit ist in vier Hauptteile gegliedert. Der erste Teil prsentiert die Grundlagen der Analyse von DNA-Methylierungsdaten sowie von Regulierungspfaden und deren funktionellen Zusammenhngen, die dabei helfen, das Phnomen der DNA-Comethylierung zu erklren. Der zweite Teil deckt das Problem des Auffindens und Korrigierens von Batch Effekten ab. Dafr wird das neu entwickelte BEclear R Softwarepacket beschrieben und mit bereits existierenden, etablierten Methoden verglichen. Zustzlich wird die optimale Strategie hinsichtlich des Umganges mit Batch Effekten, mit der diesbezglichen Parameterauswahl und mit den Auswirkungen dieser negativen Effekte auf das Endergebnis, vorgeschlagen. Der dritte Teil behandelt die Analyse von S. aureus Genomen durch die Nutzung verschiedener Techniken der Datengewinnung, die von der aufhufenden hierarchischen Clusteranalyse bis zur Affinity Propagation reichen. Interessanterweise wurden diese vieldeutigen Bakteriengenomdaten durch die gleichen exakten Matrixvervollstndigungs-Algorithmen erfolgreich komplettiert, die auch in BEclear genutzt werden. Der letzte Teil beschreibt statistische Analysen von Community aquired S. aureus Stmmen, die wichtige Verbindungen von Bakterienresistenzen und Virulenzprofilen mit der geografischen Lage aufdeckten.
Das wichtigste Ergbenis meiner Doktorarbeit ist sicherlich die tiefgrndige Begutachtung von Batch Effekten. Diese sind dafr berchtigt, den kompletten Datenverarbeitungsprozess zu beeintrchtigen. Darber hinaus knnen sie das Endergebnis ungemein beeinflussen und sind daher zur Zeit von besonderem Interesse unter Forschern, die sich mit genomweiten Hochdurchsatz- Datenstzen beschftigen. Es gibt bereits verschiedene Anstze, die es erlauben, die negativen Auswirkungen von Batch Effekten auf die erforschten Datenstze zu vermindern. Hier prsentieren wir eine neue Alternative fr diese Zielsetzung mit dem Namen "BEclear". Diese Methode erlaubt es nicht nur Batch Effekte aufzufinden und deren Auswirkungen einzuschtzen, sondern auch diese nur in den tatschlich betroffenen Teilen der Daten mithilfe von Latent Factor Model Matrixvervollstndigungs-Algorithmen zu korrigieren. Wir testeten die entwickelte Methodik an Brustkrebsdaten des The Cancer Genome Atlas Portals und verglichen die Ergebnisse mit den bereits vorhandenen Algorithmen ComBat, Surrogate Variable Analysis und Functional Normalization. Wir zeigen, dass BEclear die genannten Methoden hinsichtlich der Genauigkeit an Leistung bertrifft und gleichzeitig die Vernderung der nicht von Batch Effekten betroffenen Daten vermeidet, da es sich nur auf die beeintrchtigten Gene fokussiert. Diese Eigenschaften macht BEclear zu einem wettbewerbsfhigen Algorithmus zur Korrektur von Batch Effekten, der weithin auf DNA Methylierungsdaten oder selbst auf Genexpressionsdaten angewandt werden kann. BEclear steht als R Softwarepaket zur Verfgung.

Contents

Chapter 1

An Introduction to DNA Methylation and Cancer

Methylation is a chemical modification catalyzed by various enzymes whereby a methyl (CH_3) group is attached to specific sites of proteins, DNA and RNA. Epigenetic information is not encoded in genes but the information encoded by DNA can be directly affected by epigenetic modification such as DNA methylation [18].

One form of methylation, the most common in mammals, is the conversion of cytosine to 5-methylcytosine in the sequence of CpG dinucleotides, i.e. the cytosine base is directly followed by a guanine base. Methylation may prevent cleavage of DNA at the recognition site of a restriction enzyme. For example, the restriction enzyme Hpall cleaves CCGG, but not Cm5CGG.

The methylation reaction is catalyzed by one of several DNA methylation enzymes - DNA methyltransferase, which carries out the transfer of a methyl group from S-adenosylmethionine to cytosine. In humans and most mammals, DNA methylation is a natural DNA modification and mostly affects the base cytosine (C), facing a guanine (G), i.e. methylation occurs mainly at CpG-dinucleotides.

In differentiated cells $70-80\%$ of all CpG-dinucleotides are methylated in the human genome. However, normal tissue methylation occurs primarily in genomic regions where the density of CpG dinucleotides is low, and the majority of CpG-normal islands are completely unmethylated.

Besides all the large or small genetic variations that have been linked to many human diseases so far, we are just about to start appreciating the large amount of variability with regards to epigenetic variations in humans and between normal and disease samples [176].

1.1 DNA methylation

About 1% of all human DNA base pairs consist of methylated cytosine bases, where a methyl group is covalently attached to the C5 position of cytosine [105]. Since most of these occur in the context of CpG dinucleotides, from 60 – 90% of all CpGs are methylated in human [55]. DNA methylation is associated with parasitic DNA suppression [212], repression of gene transcription [29], and genomic imprinting [168].

In addition to that, DNA methylation plays an important role in cancer where the genome is mostly hypomethylated except for promoter regions of tumor suppressor genes that are hypermethylated [95, 111, 108].

1.2 Breast invasive carcinoma-BRCA, Kidney Renal Clear Cell Carcinoma - KIRC

1.2.1 BRCA

Mammary glands are composed of three main types of tissues - fat, connective and glandular tissue. Breast cancer (BRCA) is a malignant tumor that develops from cells namely in the glandular tissue. Breast cancer has similarities to that of other malignant tumors in the body. As a result of several mutations occurred one or more cells in the glandular tissue start abnormally fast sharing. These cells form tumors that can invade nearby tissues and create secondary tumor hearths - metastases.

Breast cancer occurs as a result of active uncontrolled division of abnormal cancer cells. Without treatment the tumor is growing rapidly in size, can grow into the skin, muscles and chest. In the lymph vessels cancer cells penetrate into the lymph nodes nearest to the breast. With the bloodstream, they spread throughout the body, giving rise to new tumors - metastasis. In most cases, breast cancer metastasizes to the lungs, liver, bone, brain. The defeat of these bodies, as well as the disintegration of the tumor, leads to death. Breast cancer can develop on the background of precancerous diseases, which include breast- and fibro-adenoma.

Genetic alterations of the $BRCA1$ and $BRCA2$ (tumor suppressor) genes have been associated with breast cancer formation [7]. Researches have discovered large number of different types of mutations of these genes [39]. Some of these are harmless, while others can cause serious issues duch as hereditary breast-ovarian cancer syndrome [1].

Identification of mutations in the genes analyzed has predictive value for determining the risk of developing breast cancer and / or ovarian cancer [8]. The discovery of a gene defect in clinically healthy women allows for timely diagnosis in case of cancer of the breast and / or ovarian cancer and warn about their possible serious consequences. For patients with confirmed malignant disease already this alarm signal makes it possible to determine its possible hereditary nature. In the studies of breast and ovarian cancers, researchers identified 10 of the most frequent mutations in the genes $BRCA1$, $BRCA2$, $CHEK2$ and $NBS1$.

The genes $BRCA1$ and $BRCA2$ (BREAST CANCER GENES 1 and 2) encode the amino acid sequence of nuclear proteins involved in the regulation of DNA repair and cell division. The intact (wild-type) forms of both genes act as a tumor suppressor and ensure the integrity of the genome. Furthermore, the protein products of the genes repress transcription of estrogen receptor, thus constraining excessive cell proliferation of breast cancer and other estrogen-bodies, in particular at puberty and pregnancy. Mutations in the genes $BRCA1$ and $BRCA2$ have been shown to lead to increased levels of chromosomal instability in cells, which may contribute to their malignant transformation. Today there are more than 1,000 different mutations in $BRCA1$ and $BRCA2$, associated with an increased risk of developing breast cancer, ovarian, prostate, colon, throat, skin, and

[1] $http://en.wikipedia.org/wiki/Hereditary\_breast\%E2\%80\%93ovarian\_cancer\_syndrome$

others. Upon detection of mutation (s) in the $BRCA1$ and $BRCA2$ genes in a woman, her individual risk of developing breast cancer and / or ovarian cancer is 50 to 80%.

1.2.2 KIRC

Kidney Renal Clear Cell Carcinoma (KIRC) accounts for most malignant kidney tumors and is known to cause fatal genitourinary diseases. Very often it is not treated by radiotherapy and chemotherapy due to inherent resistance. The metastatic phase of KIRC can currently not be cured. Thus, genome analysis can be a clue to successful diagnosis and anticipated treatment during early stages of the disease.

Research groups all over the world use advanced sequencing technologies and TCGA (more details in Subsection 1.4.2) to figure out groups of differentially expressed genes to determine subtypes of the cancer [210]. This, in turn, combined with identification of distinctly expressed genes and altered pathways is important for biomarker identification for early cancer diagnosis and treatment planning. Sophisticated computational methods can be used to identify upstream disease causal genes and assist in remedy prescription.

1.3 Role of DNA methylation in Cancer

DNA methylation is an important mechanism for regulation of gene expression. It has been shown that altered methylation patterns are associated with diseases such as various cancers, diabetes, first and second class, schizophrenia, etc. It is therefore important to be able to experimentally characterize and analyze the methylation profile of the genome.

Methylation patterns in neoplastic cells change considerably compared to normal cells, wherein the total demethylation of the genome is accompanied by an increased methyltransferase activity and hypermethylation of local CpG-islands. In all studied neoplasies such an imbalanced methylation has been observed. It is obvious that these disorders can alter the chromatin structure and function of DNA, thereby making a significant contribution to the phenotypic and genetic instability of tumor cells.

It was found that one of the primary disorders of DNA methylation in neoplastic cells, is a total genome hypomethylation. Reducing the number of methyl groups is one of the early steps, often even before the tumor develops and leads to cellular transformation. The direct role of DNA hypomethylation in cell transformation has been proved on the basis of low-methionine diet, leading to a shortage of donor of methyl groups, what causes hypomethylation of DNA and liver tumors [160].

Despite the apparent association of hypomethylation of DNA with the formation of tumors, the causes and specific mechanisms underlying its carcinogenic effect is still unclear. There is evidence that hypomethylation may affect certain oncogenes such as $KRAS$ in lung cancer and bowel in humans. These gene-specific local changes occur in the early stages of carcinogenesis, and were found in particular in benign polyps, which are precursors of colon carcinoma [82].

Disregulation of genomic imprinting as a result of demethylation and its role in carcinogenesis has been demonstrated in the study of Wilms' tumor [61]

Another consequence is a total hypomethylation resulting from perturbations of the general pattern of methylation of genomic instability. So hypomethylation of DNA in

embryonic mouse cells, gene knockout *dnmt*1, increased frequency of endogenous retroviruses and re-arrangements of parasitic sequences re-arrangements, the incidence of deletions and translocations of some unique genes, i.e. it caused chromosomal abnormalities and subsequent death [32].

Local hypermethylation is represented by a small portion (*app.*20%) of *CpG* dinucleotides , which then form *CpG*-islands. *CpG*-islands are mostly unmethylated in normal cells. Aberrant hypermethylation of *CpG*-islands is a peculiarity of immortalized and transformed cells and is associated with inactivation of certain tumor suppressor genes in human [194].

An important role in increasing local hypermethylation plays methyltransferase activity, especially as it is a characteristic feature of tumor cells [172]. In the study of some cell cultures it was shown that the increase of DNA methyltransferase activity often precedes malignant transformation. Thus, transfection of a cloned gene in human Dnmt1 immortalized human fibroblasts leading to aberrant *CpG*-island methylation in the promoter regions of several genes.

Thus, it appears that increasing Dnmt1 activity plays a role in the aberrant methylation of *CpG*-islands. However, simply increasing the level of expression can not explain the appearance in the ability of the enzyme to gain the methylation de novo. Apparently, in transformed cells and tumor, the protection mechanism is broken by *CpG*-island methylation.

Hypermethylation of *CpG*-islands results in a stable inactivation of the adjacent gene, that is the phenomenon of MAGI (methylation-associated gene inactivation). It occurs as a result of occurrence of steric hindrance to the binding of transcription factors or as a result of heterochromatin mediated binding of methylcytosine binding proteins MBD [171].

Suppression of the expression of any of the tissue-specific genes causes some damage to the differential phenotype cells, without affecting the overall viability. At the same time, the inactivation of tumor suppressor genes or gene repair can create conditions for an uncontrolled proliferation [59]. Aberrant methylation of *CpG*-islands is an early event in the process of formation of a tumor.

A characteristic feature of tumor and of transformed in vitro mammalian cells is an imbalance of methylation of genomic DNA, which makes a significant contribution to the phenotypic and genetic instability. At the same time, the instability of the $5-MeC$ composed *CpG* dinucleotide, leading to epimutations may have the same end result. Thus, methylation, as one type of the epigenetic modifications of DNA may eventually lead to genetic changes, making clear the relationship between genetic and epigenetic processes in the formation and development of tumors.

It was reported that methylation patterns appear in the early stages of malignant transformation of mammalian cells. From a medical point of view, this opens opportunities for early diagnosis and treatment of disease. Moreover, in contrast to mutations that are essentially irreversible modification of DNA, epigenetic modifications are very stable, but essentially reversible modifications.

1.4 Measuring of DNA Methylation

1.4.1 Overview of techniques

Nowadays there exist a number of different approaches for performing epigenome-wide association studies. The following methods are considered to produce accurate DNA methylation data: methylated DNA immunoprecipitation sequencing (MeDIP-*seq*), methylated DNA capture by affinity purication (MethylCap-seq), bisulphite sequencing (NGS), and Illumina arrays (Infinium HumanMethylation450 BeadChip). Depending on the goals of the study, the appropriate approach for analysis of DNA methylation must be selected. For example, NGS technologies are the gold standard to produce high resolution epigenomic data and their working procedure is relatively complicated.

Initially, genomic DNA must be purified (Phenol/Chlorophorm) and fragmented (Sonication). Further, ends are polished and methylated adaptors are added. Further, DNA strands are separated and DNA is treated with sodium bisulfit. Later DNA is purified (single stranded) and amplified by few PCR-cycles, what makes it again double stranded. Finally, library fragments are purified, materials are clustered and loaded on a sequencing device. The experimental part is followed by data processing procedures, namely process and quality control of the data to form FASTq files; aligning the reads to a reference genome (bisulfite converted), count the number of methylated and unmethylated positions. Overall, NGS results in $3 \cdot 10^{11}$ bases in 10-12 days what is equal to 4 genomes and can produce 1 TB (*fast*Q) to 40 TB (processed images) of data. In comparison to array based technologies: NGS is more complicated and less cost efficient.

Another widespread technology is represented by the Infinium HumanMethylation450 BeadChip array. It allows to interrogate more than 485 000 methylation sites per sample at single-nucleotide resolution. This array covers 99% of RefSeq genes, with an average of 17CpG sites per gene region distributed across the promoter, 5'UTR, first exon, gene body, and 3'UTR. It also delivers data for 96% of CpG islands, with additional coverage in island shores and the regions flanking them. Finally, Illumina Kit is able to run up to 96 samples in parallel what significantly reduces the time during large cohort studies. Here we analysed data produced by Infinium HumanMethylation450 that was obtained from the TCGA portal.

1.4.2 The Cancer Genome Atlas - TCGA

Cancer has hundreds of different forms depending on the organs and tissues where it originated, and on the genetic changes that cause the development of tumours and affect the outcome of treatment. Therefore, a treatment effective in one case might be useless in another one.

If it were possible to identify all the mutations that have occurred in the genomes of cancer cells of all types, and the changes caused by these mutations, and then to analyse their relationships with the course of the disease; One may be able to find molecular markers that would classify cancer cells and may select treatment in accordance with this classification. If all this were possible, humanity would have made significant progress in cancer therapy.

Such an ambitious goal was set by the creators of the project Pan-Cancer Initiative, launched in October 2012. Work on the project is a part of a cancer genome atlas (the

Cancer Genome Atlas (TCGA [2]), through which the specialists of the National Cancer Institute and the National Institute of Research of the human genome share obtained data. The results of the first phase of this work were presented in 18 articles, four of which were published in the journal Nature Genetics.

The study required the coordinated work of several dozens of research groups. Scientists analysed the results of several thousands of patients with twelve types of tumors: glioblastoma multiforme, myeloid leukemia, acute lymphoblastic, squamous cell carcinoma of the head and neck, adenocarcinoma of the lung, squamous carcinoma of the lung, breast, kidney, cervix, ovary, bladder, endometrium, colon adenocarcinoma colon and rectum. In these tumors they considered all mutations, gene copy number and the activity of their work, the metabolic pathways in which each gene is involved, the degree of DNA methylation, microRNAs and protein synthesis, and the clinical picture of the disease. Then, all the data were combined and it was tried to find a connection between them.

Malignant tumors are traditionally divided according to their place of origin, such as lung cancer, skin or intestines. As shown by molecular analysis, 'the relationship of tissue' really puts an overall mark on cancer cells, but the tumor from one organ or tissue often differ, and tumors of different organs may have common molecular markers.

Thus, the same genetic mutation may be characteristic for certain tumors and glioblastomas gastric serosa endometrial, bladder and lung. Therefore, a drug effective for lung cancer, may be appropriate for a bladder tumor with the same molecular marker too. A breast ductal cancer is a group of diseases which are controlled by different genes. Sometimes the same genetic defect causes different effects depending on the organ where it is found. Thus, a family of genes Notch, inactive in some lung tumors, head and neck, skin, cervical, is active in leukemia.

The global objective of this research of cancer is to identify biomarkers that can be used to classify tumors and to determine which treatment is best suited for each type.

To fully explore every option of cancer, the corresponding sample is subject to large-scale study using the methods of sequencing and bioinformatics: quantitative gene expression analysis, quantitative analysis of gene copy number variation, SNP genotyping, genome-wide analysis of DNA methylation patterns, sequencing of exons. The data is made publicly available, so that any researcher can see them and use them in their work. The project TCGA showed that active and large-scale cooperation of researchers from different institutions can be fruitful, and the data resulting from the work can be used by scientists around the world.

TCGA already includes samples from more than 11 000 patients for 33 types of cancer, and today is the largest collection of tumors. All these samples were analyzed for the presence of the key genomic and molecular features. Results for 13 types of cancer were published in major scientific journals. By the end of 2014 TCGA scientists had almost finished exome sequencing for most types of tumors and full genome sequencing for more than 1,000 samples of cancerous tissue. More than 2,700 scientific articles cited the TCGA portal, proving its large impact.

[2] http://cancergenome.nih.gov/

1.4.3 Illumina Infinium HumanMethylation 27, 450K Chip

DNA methylation is a crucial part in the regulation of gene expression. Cellular development and maintaining tissue identities require certain methylation levels to maintain normal functionality.

Many research projects have implicated aberrant methylation in the etiology of many human diseases, especially cancer. At the moment, several common methods for quantitative measurements of methylation exist. One of the most common is a series of microarray company Illumina. Illuminas Infinium Methylation Assay provides quantitative methylation measurements at the single-CpG-site level, offering the highest resolution for understanding epigenetic changes to date. We will focus on The HumanMethylation27 BeadChip and its successor - 450K chip.

Both of them use Infinium technology [3] described for SNP genotyping to perform genome-wide screening of DNA methylation patterns [187].

The 450K chip allows measuring the level of methylation of *CpG* sites in about 486,000, more or less evenly distributed locations throughout the genome. The technology can be described as follows. Each CpG site is measured with two fluorescent probes. The fluorescent signal is proportional to the sample according to the amount of methylated and unmethylated *CpG* sites in the test sample. The chip allows one to test up to 12 biological samples simultaneously. Thus, we have the output value table in which the number of rows equals the number of *CpG* sites, and the number of columns - the number of biological samples being analyzed.

The pipeline for data analysis using the *R* language and Bioconductor [4] libraries has approximately the following items (with the corresponding packages from Bioconductor):

- The choice of scale (β or M value [5]).
- Adjust the color balance (color channel balance adjustment). Some *CpG* sites are measured using samples of one color and some with two. This problem is eliminated by normalizing the signals of the two samples in each biological sample.
- Background correction. Each slot of biological samples on the chip has a different default background. Therefore, the alignment of values between the samples must be background corrected.
- The normalization between samples (between-sample normalization). Used are mainly quantile normalization and normalization of SVN (package *lumi* [6]).
- Testing for group effect (batch effect) using principal component analysis.
- Peak based correction.
- Correction of the effect on the group with packages *ComBat* [7] and *SVA* [8].

[3] http://www.illumina.com/technology/beadarray-technology/infinium-hd-assay.html
[4] http://www.bioconductor.org/
[5] http://www.biomedcentral.com/1471-2105/11/587
[6] http://www.bioconductor.org/packages/release/bioc/html/lumi.html
[7] jlab.byu.edu/ComBat/Abstract.html
[8] http://www.bioconductor.org/packages/release/bioc/html/sva.html

- Testing for statistical significance using linear models, permutations, or routine testing to test hypotheses (*limma* [9] packages and *multtest* [10]).
- Analysis of data using different algorithms machine learning.
- The correlation of gene expression data and SNP (methylation Quantitative Trait Loci).

[9] http://www.bioconductor.org/packages/release/bioc/html/limma.html
[10] http://www.bioconductor.org/packages/release/bioc/html/multtest.html

Chapter 2

Statistical methods

This chapter provides the methodological background for the projects that I worked on during my thesis and that are described in the following chapters.

2.1 Kolmogorov-Smirnov - KS

In statistics, there exist many criteria and statistical tests for the various types of comparisons of data. They are all based on the following main concept.

The principle steps of the analysis are:

- using experimental values the criterion is calculated according to a formula
- the experimental value is compared with the critical value (the standard set) by a certain algorithm
- a comparison is conducted for the experimentally obtained and statistically determined in the statistics critical values and conclusion on the extent of the differences of the compared data.

The Kolmogorov-Smirnov test determines whether the two compared distributions are of the same type. If we compare the experimentally obtained distribution with a normal distribution, then using this criterion enables to get an answer about whether our distribution follows the normal distribution.

The One-sided Kolmogorov-Smirnov test is based on the maximum difference between the cumulative distribution of the sample and the expected cumulative distribution:

$D_n = supp|F_n(x) - F(x)|$

With:

$F_n(x)$ the cumulative distribution of the sample;

$F(x)$ - the expected cumulative distribution (with certain parameters).

If we want to compare between two experimental distributions, it is also possible conduct this with the help of this criterion, but in this case we get a response about

whether these two belong to the distribution of any one type (binomial and Poisson, etc.) without specifying the distribution type. The principle of comparing distributions in the Kolmogorov-Smirnov test is to compare the percentile curves of the two distributions. Percentile curves are curves of the frequency distribution of data, built on the basis of summing up the accumulated frequency of all values below given.

If the Kolmogorov-Smirnov statistics D is significant, the hypothesis that the appropriate allocation is normal, must be rejected. The output probability values are based on the assumption that the mean and standard deviation of the normal distribution is known a priori and are not estimated from the data. However, in practice parameters are usually calculated directly from the data depending on how many discrete (individual) values were obtained by the test performed on the gradation values of the x-axis.

At the same time, we mark percentiles (percentiles ranks) on the y-axis. To build the curve, for each value (the value of the test results) its percentile rank is previously determined, which is obtained by adding the percentage of occurrence of this result and the percentage of occurrence of all results that lay below the given one. [53]

If two distributions are compared, then two separate corresponding percentiles curves are built (cumulative frequency). Then you need to determine the degree of divergence in between, i.e. calculate the difference between the percentile values applied to each result.

The maximum value in differences D_{max} is selected in percentiles and it becomes the experimental value for the Kolmogorov-Smirnov test [19, 37, 202].

So, suppose we have built percentiles curves for these two distributions, one of which is a normal (according to preliminary information, for example, obtained by other researchers). In order to assess the normality of our distribution, it should be compared with normal.

To this end, in relation to each category are calculated the difference between the percentile values, and the largest absolute value of D_{max} is selected. Next, we must take a critical value for this test to make a comparison of the experimental and critical values in order to build the output of differences of distributions. All statistical methods are designed to test the null hypothesis, in other words, they assess the legality of it. With regard to the Kolmogorov-Smirnov test this rule is: if the experimental value of the test is equal to or greater than the critical value, the hypothesis of significant differences is accepted.

Limitations of the Kolmogorov-Smirnov test

The criterion requires that the sample was large enough. When comparing the two empirical distributions $n_{1,2} \geq 50$ is needed. Comparison of the empirical distribution with the theoretical sometimes is allowed in cases of $n \geq 5$ [75].

Categories should be arranged in ascending or descending of some entity. For example, you can take as a category day of the week, or the 1st, 2nd, 3rd months after a course of therapy, increased body temperature, increased feelings of failure, etc. At the same time, if we take the level which happened to be aligned to this sequence, then it will be impossible to accumulate the frequency of categories, because they differ only in quality and do not represent the scale of the order.

2.2 False Discovery Rate p-value adjustment

False discovery rate (FDR) p-value adjustment algorithm was proposed by Benjamini and Hochberg in 1995 [16] and is used for controlling the expected rate of rejected null hypotheses which appeared to be false discoveries when a large number of tests is being conducted. In comparison to its main competing method, which is familywise error rate (FWER), FDR is more flexible and allows a small number of tests to be wrongly detected by providing not so rigorous regulation of type I errors (false positives). The problem is formulated as following. Assume the simultaneous testing of m null hypothesis is performed and m_0 is the number of true null hypothesis. We will use the common variables:

- V number of false positives;
- S true positives;
- T false negatives;
- U true negatives;
- R number of rejected null hypotheses.

For these variables, the next equations are valid:

- $R = V + S$
- $m - R = U + T$
- $m_0 = V + U$
- $m - m_0 = S + T$

Now the rate of erroneously rejected null hypotheses can be defined as $Q = \frac{V}{V+S}$. Certainly, when $V + S = 0$ this rate Q will be also equal to 0. Finally, the false discovery rate FDR is defined as Q_e and is the expectation of Q. This expectation can be formulated as:

$$FDR = E[Q] = E[\frac{V}{V+S}] = E[\frac{V}{R}].$$

Thus the procedure of FDR correction efficiently finds a certain threshold, which is used to define tests to be significant or no at the level q. As a result, the equation defined above implies that all procedures that control the familywise error rate also control the false discovery rate. But if some procedure controls exclusively FDR, than the gain in power is assumed [16]. For example, if the number of non-true null hypothesis is large, then the number of true positives S is abundant as well. This leads to the increase of difference between error rates. Y. Benjamini and Y. Hochberg defined the false discovery rate procedure as following (cited from [16]).

"Consider testing $H_1, H_2, ..., H_m$ based on the corresponding p-values $P_1, P_2, ..., P_m$. Let $P_{(l)} \leq P_{(2)} \leq ... < P_{(m)}$ be the ordered p-values, and denote by $H_{(i)}$ the null hypothesis corresponding to $P_{(i)}$. Define the following Bonferroni-type multiple-testing procedure:

let k be the largest i for which $P_{(i)} \leq \frac{i}{m}q^*$;

then reject all $H_{(i)} i = 1, 2, ..., k$. (1)

Theorem 1. For independent test statistics and for any configuration of false null hypotheses, the above procedure controls the FDR at q^*.
Lemma. For any $0 \leq m_0 \leq m$ independent p-values corresponding to true null hypotheses, and for any values that the $m_1 = m - m_O$ p-values corresponding to the false null hypotheses can take, the multiple-testing procedure defined by procedure (1) above satisfies the inequality

$$E(\mathbf{Q}|P_{m_0+1} = p_1, ..., p_m = p_{m_1}) \leq \frac{i}{m} q^*. \quad (2)$$

Now, suppose that $m_1 = m - m_O$ of the hypotheses are false. Whatever the joint distribution of P_1", ..., P_{m_1}" which corresponds to these false hypotheses is, integration inequality (2) above we obtain

$$E(\mathbf{Q}) \leq \frac{i}{m} q^* \leq q^*,$$

and the FDR is controlled.

Remark. Note that the independence of the test statistics corresponding to the false null hypotheses is not needed for the proof of the theorem.

This procedure was mentioned by Simes (1986) as an exploratory extension to his procedure for rejecting the intersection hypotheses that all null hypotheses are true if, for some i, $P_{(i)} \leq \frac{i\alpha}{m}$. Whereas Simes (1986) showed that his procedure controls the FWER under the intersection null hypothesis, Hommel (1988) showed that the extended procedure for inference on individual hypotheses does not control the FWER in the strong sense: for some configuration of the false null hypotheses, the probability of an erroneous rejection is greater than α. Hochberg (1988) has suggested a different way to utilize Simes's procedure so that it controls the FWER in the strong sense, by offering the following procedure:

let k be the largest i for which $P_{(i)} \leq \frac{i}{m+1-i}\alpha$;
then reject all $H_{(i)} i = 1, 2, ..., k$.
Note the relationship between Hochberg's procedure and the FDR controlling procedure when q^* is chosen to equal α. Both Hochberg's procedure and the FDR controlling procedure are step-down procedures, which start by comparing $p_{(m)}$ with α, and if smaller all hypotheses are rejected - as if a per comparison error rate (PCER) approach had been taken. If $p_{(m)} > \alpha$ one should proceed to smaller p-values until it satisfies the condition. The procedures end, if not terminated earlier, by comparing $p_{(1)}$ with α/m, as in a pure Bonferroni comparison. At the two ends the procedures are similar, but, in between, the sequence of $p_{(i)}$s is compared with $\frac{1-(i-1)}{m}\alpha$ in the current procedure, rather than with $\frac{1}{m+1-i}\alpha$ in Hochberg's procedure. The series of linearly decreasing constants of the FDR controlling method is always larger than the hyperbolically decreasing constrants of Hochberg, and the extreme ratio is as large as $\frac{4m}{(m+1)^2}$ at $i = \frac{m+1}{2}$. This shows that the suggested procedure rejects samplewise at least as many hypotheses as Hochberg's method and therefore has also greater power than other FWER controlling methods such as Holm's (1979)."

2.3 Latent Factor Models - LFM

The two primary areas of collaborative filtering are the neighborhood methods and latent factor models. Neighborhood methods are centered on computing the relationships between items. Item-oriented approach evaluates a preference for an item based on ratings of neighboring items by the same user. In practice this methods are applied, for example, to predict the rating of a user for some item (film or item on the market). The neighbors of a product are other products that tend to get similar ratings when rated by the same user.

Very often in practice data sets are not complete, i.e. they contain missing values. In Bionformatics, scientists encounter these cases quite frequently. We will not be covering the causes of incomplete data, yet, would adopt methods to overcome this issue. *Latent factor models* applied to identify user's preferences for a movie can be tailored and used here to supplement the missing entries in data set.

The straightforward way to overcome this problem is simply a removal of rows/-columns. However this may lead to considerable data distortion which is not acceptable for comprehensive research.

Latent factor models are an alternative approach that tries to explain the ratings by characterizing both items and users based on the inferred from the ratings patterns. Thus, these patterns can be used to extract useful information about genes as well.

Latent Factor Models (LFM) - is a novel technique which can be applied to data sets with missing values. Some of the most well-known and effective realizations of latent factor models are based on *matrix factorization*. In its basic form, matrix factorization characterizes both items and users by vectors of factors inferred from item rating patterns. High correspondence between between these two factors leads to a similarity. These methods have become popular in recent years by combining good scalability with predictive accuracy. In addition, they offer much flexibility for modelling various real-life situations.

One strength of matrix factorization is that it allows incorporation of additional information. Recommender systems rely on different types of input data, which are often placed in a matrix with one dimension representing users and the other dimension representing items of interest. Thus, formally the problem is follows: we want to recover matrix D (size $m \times n$), but have access to only k of its entries, where k is much smaller than the total number of entries (i.e. $m \cdot n$) In general, it seems to be impossible without some extra information.

Suppose we would like to recover a square $n \times n$ matrix D of rank r. Although D contains n^2 entries, our assumption of its rank r means that it can be represented exactly using singular value decomposition (SVD) [71].

$$D = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

With: V^T is transposed of V. Σ is an $r \times r$ diagonal matrix with real, positive elements $\sigma_i > 0$. U is an $n \times r$ matrix with orthonormal columns $u_1, ..., u_r$. That is, $u_i \cdot u_i^T = 1$ and $u_i^T \cdot u_j = 0$ if $i \neq j$. V is also $n \times r$ with orthonormal columns $v_1, ..., v_r$. The column space of D is spanned by the columns of U, and the row space is spanned by the columns of$V$.

The number of degrees of freedom associated with a rank r matrix D is $r(2nr)$. To see this, note that Σ has r non-zero entries, and U and V each have nr total entries. Since U and V each satisfy $r(r+1)/2$ orthogonality constraints, the total number of degrees of freedom is $r+2nr-r(r+1) = r(2nr)$. Thus, when r is much smaller than n, there are significantly fewer degrees of freedom than the size of D would suggest. The question is then whether D can be recovered from a suitably chosen sampling of its entries without collecting the double amount of measurements. The authors of [27] consider an alternative which minimizes the sum of the singular values over the constraint set. This sum is called the nuclear norm.

In our approach we, analogously to [103], incorporate Gradient descent optimization algorithm (see chapter 4). The algorithm loops through all ratings in the training set. For each given training case, the system predicts d_{mn} and computes the associated prediction error:

$$Error_{ij} = D_{ij} - L_i^T \cdot R_j$$

Then it modifies the parameters by a magnitude proportional to γ in the opposite direction of the gradient.

$$L_i \leftarrow L_i + \gamma \cdot (Error_{ij} \cdot R_j - \alpha L_i)$$

$$R_j \leftarrow R_j + \gamma \cdot (Error_{ij} \cdot L_i - \alpha R_j)$$

The main advantage of this method is the ability to incorporate both gene and sample preferences by taking into account the values of neighboring entries when predicting a missing value. Assuming that the data set is represented by the $m \times n$ matrix D, given rank r, LFM iteratively constructs an $m \times r$ matrix L and an $r \times n$ matrix R such that matrix multiplication $[LR]_{ij}$ approximately equals to D_{ij} for every unaffected entry A. The gradient descent optimization algorithm was used to minimize the global loss i.e. the difference between $[LR]$ and D [103].

We explain in detail the usage of LFM (with corresponding α and γ parameters) in Section 4.

To sum up, the system learns the model by fitting the previously observed entries. However, the goal is to generalize those previous entries in a way that predicts future, unknown entries. Thus, the system should avoid overfitting the observed data by regularizing the learned parameters, whose magnitudes are penalized.

Stochastic gradient descent

There are two approaches to minimizing an error. These are *stochastic gradient descent* and alternating least squares *(ALS)*.

The first algorithm loops through all entries in the training set. For each given training case, the system predicts r_{ui} and computes the associated prediction error. Then it modifies the parameters by a magnitude proportional to gradient in the opposite direction. This popular optimization algorithm combines implementation ease with a relatively fast running time. Yet, in some cases, it is beneficial to use ALS optimization. While in general stochastic gradient descent is easier and faster than ALS, ALS has at least two main benefits. Firstly, it allows the usage of parallelization. In ALS, the system computes each L_i independently of the other item factors and computes each R_j independently of the other user factors. This gives rise to potentially massive parallelization of the algorithm. Secondly, for systems centered on implicit data. Because the training set cannot be considered sparse, looping over each single training case, as gradient

descent does, would not be practical. ALS can efficiently handle such cases [89].

One of the main advantages of the matrix factorization approach to collaborative filtering is its flexibility in dealing with various data aspects and other application-specific requirements. A lot of the observed variation in values within the studied array is due to effects associated with *biases* (inclination towards something, or a predisposition, partiality, prejudice, preference, or predilection)or *intercepts*, independent of any interactions. For instance, common collaborative filtering data exhibits large systematic tendencies for some users to give higher ratings than others, and for some items to receive higher ratings than others. After all, some products are widely perceived as better (or worse) than others. Thus, the same computational approach is useful when working with data sets of genetic information.

2.4 Principal component analysis

Principal component analysis is one of the main ways to reduce the dimensionality of the data, having lost the least amount of information. Principal component analysis (PCA) has many diverse interpretations. The basic one is a projection method which finds projections of maximal variability. It searches for linear combinations of the columns of X with the largest or the smallest variance. Since the variance can be changed by rescaling the combination, the combinations are modified so that they have unit length (what is valid for projections).

Let S stand for the covariance matrix and X for the original data. Then it is defined by the following equation:

$nS = (Xn^1 11^T - X)^T (Xn^1 11^T X) = (X^T Xn\overline{x}\overline{x}^T)$

where $\overline{x} = 1^T X/n$ is the row vector of means of the variables. Then xa stands for the sample variance of a linear combination of a row vector x, and is equal to the $a^T \Sigma a$. It should be maximized or minimized with the subject to $\|a\|^2 = a^T a = 1$. Due to the fact that Σ is a non-negative matrix, eigendecomposition can be applied.

$\Sigma = CT\Lambda C$

where Λ represents the diagonal matrix of eigenvalues in descending order. Let $b = Ca$. b in this case is of the same length as a because of orthogonality of C. Then the problem can be modified to maximizing $b^T \Lambda b = \lambda_i b_i^2$ with the subject to $\Sigma b_i^2 = 1$.

If b is the first unit vector, then the variance is maximized. Alternatively, the same can be achieved if a is the column eigenvector corresponding to the largest eigenvalue of Σ. Considering subsequent eigenvectors yields in covering the largest combined variance whereas the chosen linear combinations are not correlated with each other. As a result the ith principal component is indeed the ith linear combination selected by the algorithm.

k initial principal components represent the best k- dimensional projection of the data. The covariance matrix is maximized whereas the sum of squared distances between the original points and their projections is minimal yielding the best approximation. First several principal components (PCs) are usually considered when searching for specific associations or patterns in the data (generally, 2 PCs are enough for plotting and for deeper analysis the number of PCs representing 90% of variance is considered as a rule of thumb). It should be pointed, that principal components are highly dependent on

the scaling of the initial data. This should be avoided unless the original variables are of similar units. Alternatively, principle components of the correlation matrix should be computed while all the initial variables should be scaled to have unit sample variance.

PCA is the approach to the assessment of the main component of the variance proportion contained in the data. It implicitly assumes that there is no separation of the real signal and technical noise. So, other heuristics are often more productive, based on the hypothesis of a "signal" (a relatively small dimension, a relatively large amplitude) and "noise" (large dimension, a relatively small amplitude). From this perspective the principal component analysis works as a filter: the signal contained mainly in the projection onto the first principal component and the remaining components of the much higher proportion of the noise.

One ancillary use of principal component analysis is to sphere the data. After transformation to principal components, the coordinates are uncorrelated, but they now have different levels of variance. Sphering the data needs to modify the scale for every principal component so that they will have unit variance. It also results in the change of variance matrix to become identity matrix. In case when samples are represented by the points following normal distribution, the point cloud would look spherical, and many measures of interestingness in exploratory projection pursuit look for features in sphered data. Borrowing a term from time series, sphering is sometimes known as *pre-whitening* [170].

The search for principal components can be cut down to the performing of singular value decomposition of the given data matrix, or to the computing of eigenvalues and eigenvectors of the covariance matrix (obtained from the initial data matrix).

The task of analyzing the main component has at least four basic versions:

- approximate data linear manifolds of smaller dimension;
- find a subspace of smaller dimension, in an orthogonal projection on which data spread (i. e. the standard deviation from the average value) is maximal;
- find a subspace of smaller dimension, in an orthogonal projection on which the mean square distance between the point of maximum is projected;
- for this multi-dimensional random variable to construct an orthogonal coordinate transformation, in which the correlations between the individual coordinates vanish.

The first three options operate with finite sets of data. They are equivalent, and are not using a statistical hypothesis about the generation of data. The fourth option operates with random variables. The principal component analysis is always applicable. The usual assumption about the applicability to the normally distributed data (or the data with the distribution close to normal) is wrong. However, the method is not always effective in reducing the dimension with the given constraints on accuracy. For example, data with high accuracy can follow any curve, and the curve may be differently located in the data space. In this case, the principal component analysis for the acceptable accuracy requires several components (instead of one), or not at all will reduce the dimension with acceptable accuracy.

The principal component analysis is heavily used in Bioinformatics to reduce the dimensionality of description, highlight important information, data visualization, and others.

2.5 Hierarchical clustering

Hierarchical clustering is a set of algorithms for organizing data visualization which is provided via graphs.

Algorithms for ordering this type of data are based on the fact that a certain set of objects are characterized by a certain degree of coherence. They presuppose the existence of sub-groups (clusters of different order). Algorithms, in turn, are subdivided into agglomerate (Unity) and devising (sharing). Based on the number of treats they are sometimes devided into emit monothetic and polythetic classification methods. Like most visual ways of presenting dependency graphs they quickly lose visibility by increasing the number of objects. There are a number of specialized programs for the construction of graphs.

Devising clustering. First, each object is considered a separate cluster. Singleton clusters are naturally determined by the distance function:

$R(\{x\}, \{x'\}) = p(x, x')$

Then the algorithm starts the process of merges. At each iteration, the pair of the most close clusters U and V form a new cluster $W = U \cup V$. The distance from the new cluster W to any other cluster S is calculated by the distances

$R(U \cup V, S) = \alpha_U R(U, S) + \alpha_U R(V, S) + \beta R(U, V) + \gamma |R(U, S) - R(V, S)|$

This universal formula generalizes practically all reasonable means to determine the distance between the clusters. It was presented by Lance and Williams and in 1967. [113].

The most time consuming operation in Algorithm is the search for the nearest pair of clusters. It requires $O(l^2)$ operations within the main loop. Accordingly, the construction of all taxonomic trees requires $O(l^3)$ operations. This limits the applicability to the samples of the length of a few hundred objects. The idea of acceleration algorithm is to sort out only the closest couples. A parameter δ is set and only those pairs are computed which convey the rule:

$\{(U, V) : R(U, V) \leq \delta\}$ When all pairs have been exhausted, the parameter δ increases, and formed a new set of pairs. And so on, until complete fusion of all objects in one cluster is obtained.

A dendrogram usually denotes a tree , i. e. a graph without cycles, built on a matrix of proximity measures. Dendrogram represents a mutual communication between objects in a given set. A similarity matrix (or differences) is required to create a dendrogram, which determines the level of similarity between pairs of objects. Most used methods are agglomerative methods. Next, you must choose a method of constructing a dendrogram, which determines how the conversion of the matrix of similarity (difference) after the merger (or division) joins next two objects in a cluster.

Dendrogram allows representing a cluster structure in the form of a flat schedule regardless of the dimensions of the original item space. There are other ways to visualize multidimensional data such as a multidimensional scaling or Kohonen maps, but they bring into the picture artificially used distortions, the effect of which is quite difficult to assess.

Hierarchical clustering produces a set of clusters, usually one with k clusters for each $k = n, ..., 2$, successively amalgamating groups. The main differences are in calculating

group-group dissimilarities from point-point dissimilarities. Many methods are based on a measure of the similarity or dissimilarity between cases, but some need the data matrix itself. A dissimilarity coefficient d is symmetric ($d(A, B) = d(B, A)$), non-negative and $d(A, A)$ is zero. A similarity coefficient has the scale reversed.

Among the hierarchical clustering algorithms are two main types: ascending and descending algorithms. Decreasing algorithms operate on the principle of "top-down". In the beginning all the objects placed in a single cluster, which is then broken down into smaller clusters. More common ascending algorithms that are placed at the beginning of each object in a separate cluster and the clusters are then combined into larger and larger, until all objects sample will not be contained in a single cluster. This results into a system of nested partitions. The results of these algorithms are usually in the form of a tree - a dendrogram. A classic example of such a tree is the classification of animals and plants. To calculate the distance between the clusters often two distances are used: a single link or a complete link. One of the disadvantages of hierarchical algorithms is that full system partitions can be included, which may be unnecessary in the context of the problem being solved. Issues when solving clustering tasks are:

- The solution of the problem of clustering is fundamentally ambiguous:
- there is no exact formulation of the problem of clustering;
- There are many quality criteria in clustering;
- There are many heuristic clustering techniques ;
- the number of clusters is usually unknown in advance;
- clustering result depends essentially on the metric that is defined by the expert in a subjective manner.

Clustering or natural classification is the process of combining objects into groups with similar characteristics. In contrast to conventional classification, where the number of groups of objects is fixed and predetermined set of ideals, the group is not pre-defined and generated during the operation of the system on the basis of a particular measure of the proximity of objects.

2.6 Affinity propagation clustering

As described before, the standard approach for clustering is to distribute the data into groups (in other words clusters) by defining centers so that the sum of squared errors between members of clusters and their central points is as small as possible. In case when real data points serve as these cluster centers, they are named exemplars. Consequently we can say that exemplars are data points, which represent complete clusters. Affinity propagation is a clustering technique which is able to define exemplars and their corresponding clusters efficiently [67]. This concept further helps in explaining the main advantage of Affinity propagation over its main concurrent, widely known clustering approach k-centers clustering. k-centers clustering starts with randomly chosen exemplars and constantly adjusts this set in order to lower the sum of squared errors. As a result, this method is very dependent on the initial set of exemplars and it should be performed several times to get the proper clustering result. But this approach fails when the number of clusters is large and the chance of generating a proper initial set of exemplars is

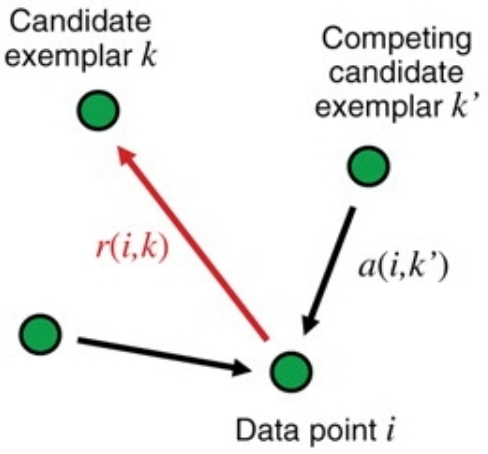


FIGURE 2.1: Sending responsibilities $r(i, k)$.

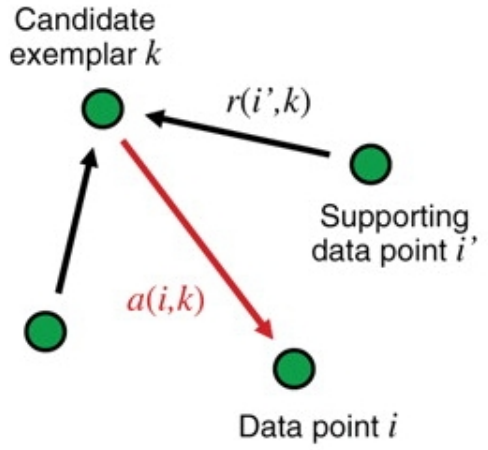


FIGURE 2.2: Sending availabilities $a(i, k)$.

low. In contrary, Affinity propagation uses another approach. It assumes all data points as prospective exemplars by considering them as network nodes. And iteratively sending real-valued messages through network edges if finds the proper clustering of data as well as exemplars which represent those clusters. The following types of messages are used in this procedure:

Issues when solving clustering tasks:

- Responsibilities $r(i, k)$. These messages are sent from data points to candidate exemplars and indicate how strongly each data point favors the candidate exemplar over other candidate exemplar (Fig 2.1).
- Availabilities $a(i, k)$. This type of messages is transmitted in the opposite direction from possible exemplars towards data points. They show the level of availability of every possible exemplar as a cluster center for every specific data point (Fig 2.2).

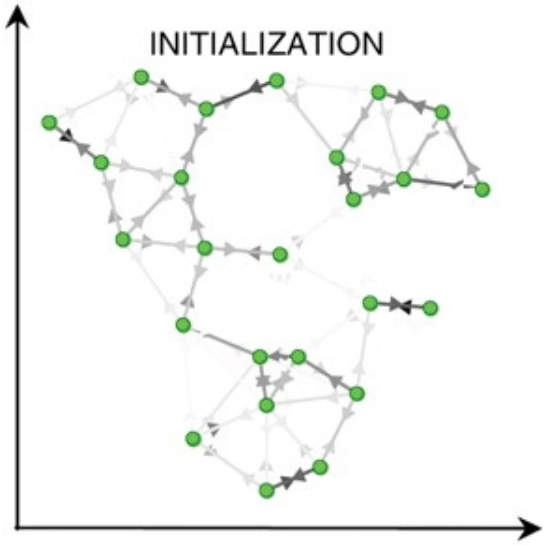


FIGURE 2.3: Messages are sent between the data points. The darkness of the arrow directed from point i to point k corresponds to the strength of the transmitted message that point i belongs to exemplar point k.

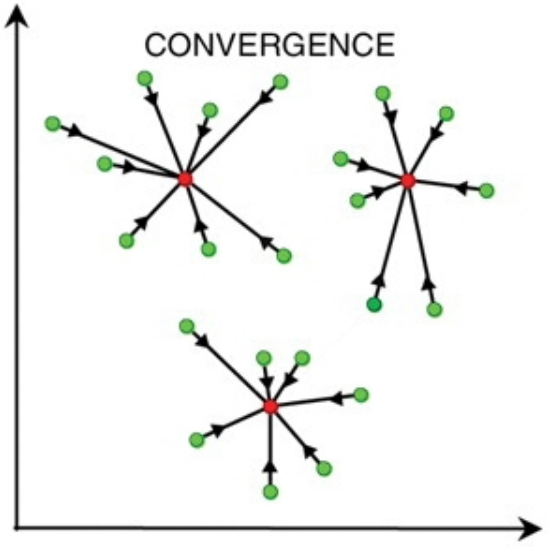


FIGURE 2.4: The result of the Affinity Propagation technique. The data is divided into three clusters. Exemplars (red data points) represent clusters and they are located in the centers. Other cluster members are marked green.

Transaction of real-valued messages throughout all data points is shown on Fig 2.3. This process ends when the selection of exemplars and corresponding clusters reaches an acceptable sum of squared errors (Fig 2.4).

The main steps of the Affinity Propagation Algorithm are:

1. Input data:

1.1. Compute similarities of the data points $s(i,k)$. $s(i,k) = -\|x_i - x_k\|^2$

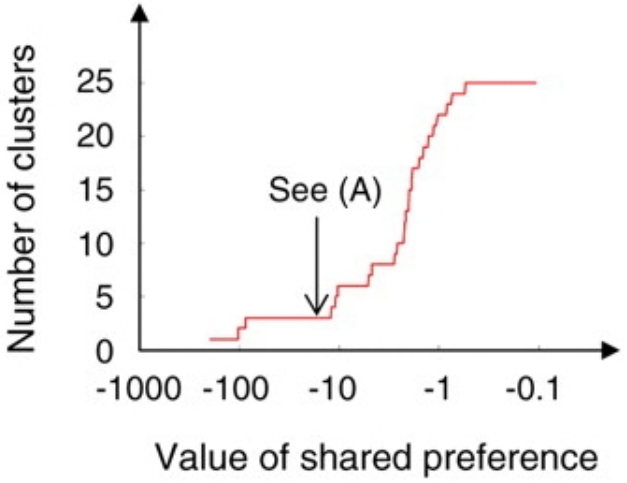


FIGURE 2.5: The effect of the value of the input preference on the number of identified exemplars.

1.2. Determine preferences $p(i)$ (or alternatively $s(i,i)$) of the data points. Higher $p(i)$ means, that these data points are more likely to be chosen as exemplars. In case when all data points have similar initial preference to serve as exemplars, then the value for this preference is selected depending on the number of clusters (Fig 2.4).

2. Initialization. $a(i,k) = 0$.

3. Compute messages.
$r(i,k) \longleftarrow s(i,k) - \max_{k' s.t. k' \neq k} (a(i,k') + s(i,k'))$

$a(i,k) \longleftarrow \min\{0, r(k,k) + \sum_{i' s.t. i' \notin i,k} max(0, r(i',k)\}$

4. Update messages.

4.1. Self-availability $a(k,k)$ is updated as follows:
$a(k,k) \longleftarrow \sum_{k' s.t. k' \neq k} max\{0, r(i',k)\}$
4.2. Update messages according to the following rule:
$mes_{new} = \lambda \cdot mes_{prev} + (1-\lambda) \cdot mes_{upd}$
where λ is a damping factor, which is between 0 and 1. By default $\lambda = 0.5$. mes_{new} - new message, mes_{prev} previous message and mes_{upd} update value for the messages. This is calculated for both message types (responsibility and availability).

5. If the number of iteration is less than the initially fixed number or data changes in the messages are above the initially fixed threshold (or they do not remain the same during some number of iterations) go to step 3, otherwise stop.

Brendan J. Frey and Delbert Dueck have used Affinity Propagation to solve a variety of clustering problems, among them are: clustering faces, detecting genes, identifying key

sentences and air-travel routing [67]. These authors discovered two main advantages of Affinity propagation (AP) over k-centers clustering: the computational time needed for AP was approximately one-hundred time less and AP assigned data points to the cluster more accurately yielding small misclassification error rate. In conclusion, the following main advantages of the Affinity Propagation clustering should be mentioned:

- AP is easy to implement;
- This clustering method outperforms commonly used k-centers clustering;
- It is able to solve random satisfiability problems where the size grows with the order of magnitude speed;
- AP can solve the NP-hard two dimensional phase unwrapping problem;
- When analyzing stereo images, Affinity propagation can successfully evaluate the depth.

Chapter 3

DNA co-methylation analysis

This chapter is based on the publication entitled "DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples" by Akulenko et al. in journal Human molecular genetics (2013) [3].

Localized promoter hypermethylation and overall DNA hypomethylation have been associated with the presence of tumor in human. Yet, despite the large amount of recently produced epigenetic data, there is still a lack of understanding how several genes behave in tumor cells with respect to their epigenetic alterations such as DNA methylation.

In this chapter we performed a novel type of analysis that measures the correlation of the DNA methylation levels of two genes across many samples. We linked this so-called co-methylation to the genomic distance of the genes, their functional similarity, and their expression levels.

Besides all the large or small genetic variations that have been linked to many human diseases so far, we are just about to start appreciating the large amount of variability with regards to epigenetic variations in humans and between normal and disease samples. Clearly, the epigenome is different and it impacts gene expression [176]. Epigenetic information is not encoded in genes but the information encoded by DNA can be directly affected by epigenetic modification such as DNA methylation [18].

About 1% of all human DNA base pairs consist of methylated cytosine bases, where a methyl group is covalently attached to the C5 position of cytosine [105]. Since most of these occur in the context of CpG dinucleotides, $60 - 90\%$ of all CpGs are methylated in human [55]. DNA methylation is associated with parasitic DNA suppression [212], repression of gene transcription [29], and genomic imprinting [168]. In addition to that, DNA methylation plays an important role in cancer where the genome is mostly hypomethylated except for promoter regions of tumor suppressor genes that are hypermethylated [95, 111, 108].

In our analysis we focused on breast cancer, a genetically heterogeneous type of cancer that belongs to the most prevalent and best studied ones [146]. The OMIM database contains 22 genes, mutations of which are associated with this cancer type (Online Mendelian Inheritance in Man, *OMIM*®. John Hopkins University, Baltimore, MD. MIM Number: 114480. [1]). Among these, the important BRCA1 gene has been shown to show cancer-specific methylation patterns [28] and a number of other genes such

[1] http://omim.org

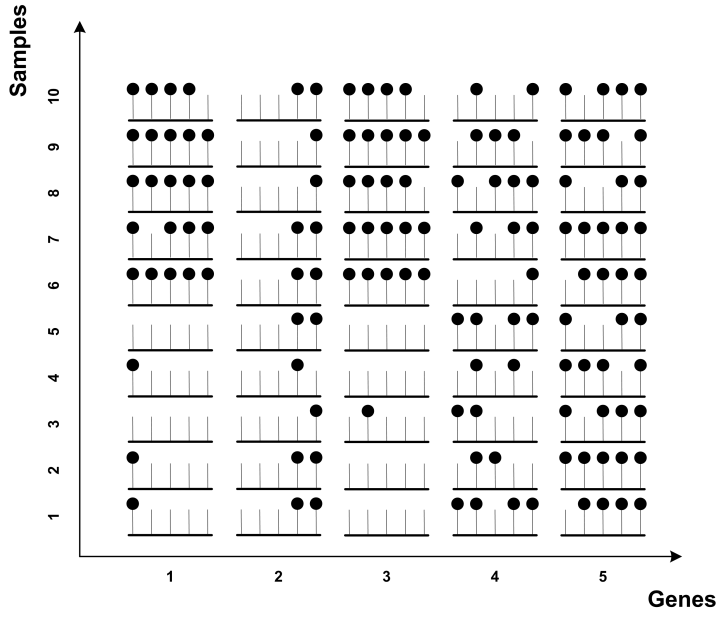


FIGURE 3.1: Schematic example of CpG methylation in five genes. The ticks indicate CpG sites. Filled circles indicate CpG methylation. The first and the third gene show highly correlated methylation levels across the 10 samples. Here, we term this behavior 'co-methylation'. The second gene is mostly unmethylated. Even though genes five and four are mostly methylated, they are not co-methylated.

as RASSF1, ARHGDIB, GRB7, SEMA3B, MMP7, PEG10, GSTP1, CHI3L2 [84]. Thus, there exists ample evidence that altered DNA methylation is associated with the development of breast cancer [211] .

Since huge amounts of epigenetic data are nowadays being generated thanks to modern technologies such as ChIP-on-chip, ChIP-seq, and bisulfite sequencing, the new field of computational epigenetics aims to analyze these data and to link them to functional data [20]. A number of international projects like TCGA, BLUEPRINT, DEEP, AHEAD, ENCODE, HEP and IHEC initiated human epigenome sequencing and mapping. A recent study reported an association of the promoter methylation profile of P14ARF, MDM2, TP53 and PTEN genes with regulatory pathways of the tumor suppressor gene TP53 [14]. Generally, functional similarity or participation in a common pathway is known to lead to gene co-expression [87, 12, 207]. This motivated us to investigate in an analogous way the possibility of co-methylation of genes across samples. Fig. 3.1 illustrates the main idea behind this study.

We based our analysis on breast cancer samples from the TCGA initiative that collects and analyzes tumor and non-tumor samples and makes it available to the public through their data portal. We analyzed possible relations between DNA-co-methylation and genomic distance, functional similarity or pathway enrichment. We cover in detail all used data set and its processing routins in the Section 3.1.

3.1 Materials

Tumor data

DNA methylation data for tissue samples from breast invasive carcinoma patients were obtained from The Cancer Genome Atlas (TCGA) Data Portal [2]. The level 3 DNA methylation data which had been used was deposited by the group of Dr. Peter Laird of Johns Hopkins and University of Southern California (USC) [3] and consists of preprocessed DNA methylation data that was obtained using Illuminas Infinium Human DNA Methylation 27 platform. This BeadChip detects 27,578 CpG Sites in 14,475 RefSeq Genes, among which are 982 cancer-related targets. Since the data is deposited continuously, we first analyzed 183 available tumor samples deposited in September 2011 (tumor group 1) and then an additional 134 tumor samples (tumor group 2) as well as 27 matched tissue mostly from adjacent normal tissues that were both deposited in October 2011.

3.1.1 Randomization of data

Comparison against randomized data

In order to characterize the statistical significance of the correlations of gene methylation, we repeated the same steps with random data. For this, we generated a random permutation on a per gene level of the original data using the 'sample' function in R so that the distribution of the $\beta - value$ in the randomly permutated data is identical to the distribution of $\beta - value$ in the real data.

The algorithm works as follows: all $\beta - values$ are selected for the specific gene among all samples. As a result we have 344 entries for one gene which are then randomly permutated by the 'sample' function. Afterwards, permutated $\beta - values$ are assigned back to random samples that are mostly different from the original samples. Effectively, this changes their order of appearance during the correlation computing. This procedure is repeated for all 13313 genes.

3.1.2 Detailed description of data base set up

Preprocessing of raw data

Every sample is represented by a separate *.txt* file that contains the barcode of the samples, the $\beta - value$ between 0 and 1 (ratio of methylated to the sum of methylated and unmethylated sites), gene symbol, chromosome and gene position. Using Microsoft SQL Server 2008 Express all 344 samples were uploaded to two tables of our database and parsed. About 73 genes that contained NA $\beta - values$ or did not contain gene symbols were removed. If there were several entries for the same gene within a single sample, the average $\beta - value$ was computed and assigned to that gene so that every gene had only one respective $\beta - value$. Thereby the number of entries for each sample decreased from approximately 27500 to 13313.

Data filtering

The group of Dr. Peter Laird kindly made available to us a list of 2676 bad probes that they had identified in the deposited raw Illumina27k data, which were apparently affected by batch effects. Thus, in the first filtering stage we excluded any pair of genes showing correlated methylation in the combined cancer and normal samples if at least one of the two genes belonged to the list of bad probes Results reported on Figure 3.2.

[2] http://cancergenome.nih.gov/
[3] http://www.usc.edu/

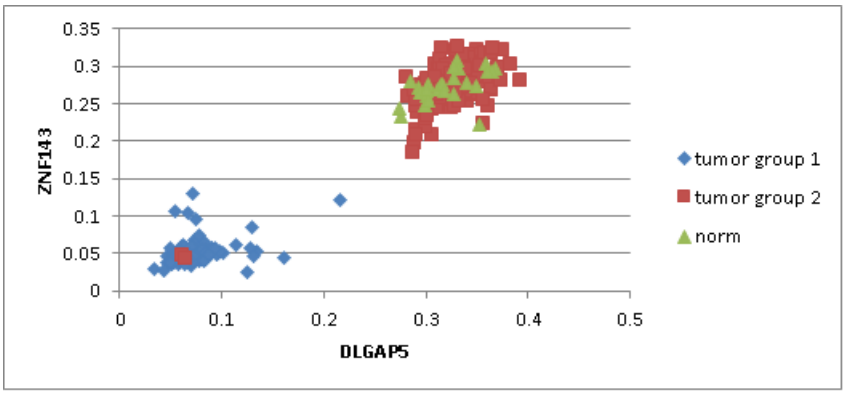


FIGURE 3.2: Examples of three undesirable cases where nave data processing indicates a high correlation of co-methylation. All such cases were removed by applying suitable filtering steps described in the data filtering section. In all three cases, the Pearson correlation of pairs of genes is $r = 0.98$. A. ZNF143 and DLGAP5 belong to the top 10 genes that were most affected by the batch effect. Samples from tumor groups 1 and 2 show significantly different methylation levels whereas samples from tumor group 2 and normal samples overlap. These associations were removed during stage one of filtering.

To avoid cases where significant correlation is found due to single very high or very low $\beta-values$ as in the case of the genes CLK1 and YPF5 (Figure 3.3), only genes were kept after stage two filtering that have no outlier $\beta - values$ according to the 'boxplot.stats' function in R.

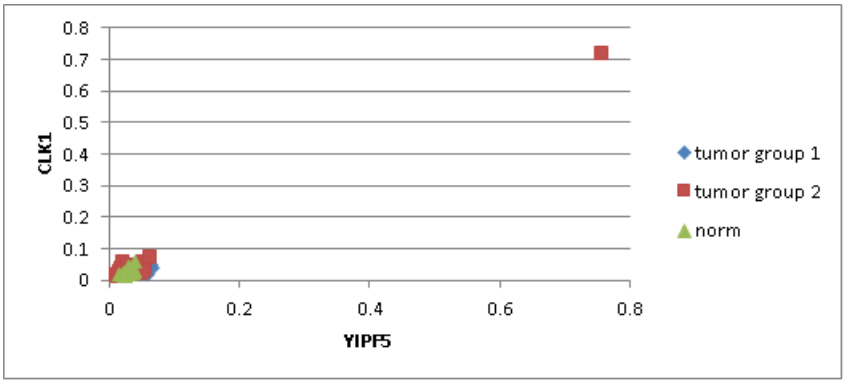


FIGURE 3.3: Examples of three undesirable cases where naive data processing indicates a high correlation of co-methylation. All such cases were removed by applying suitable filtering steps described in the data filtering section. In all three cases, the Pearson correlation of pairs of genes is r = 0.98. The two genes CLK1 and YIPF5 have the same level of methylation and both genes were not affected by batch effect. Also, the difference between maximal and minimal $\beta - values$ is high due to one outlier data point. To avoid such cases showing very small variation of $\beta - values$, such pairs were filtered out at stage two filtering.

The third stage of filtering aimed at removing genes with very small variance of the $\beta - values$ like in the case of the two genes C1R and LEMD3 (Figure 3.4). For this, we required that the third and first quartiles, respectively, of all $\beta - values$ for a single gene differed by more than 0.1.

Functional similarity

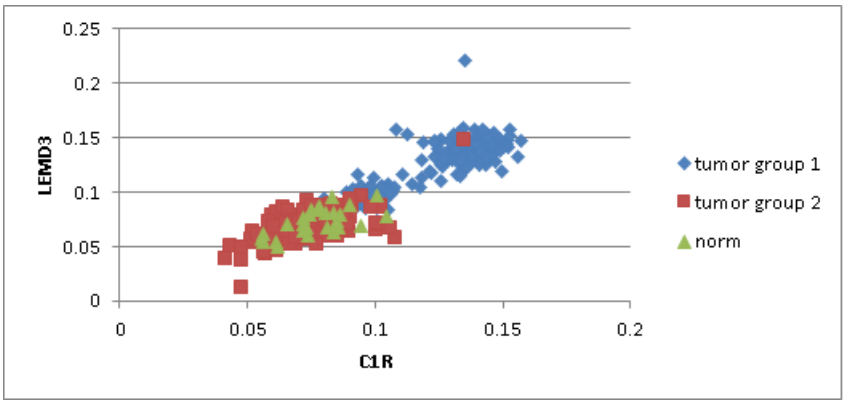


FIGURE 3.4: Due to three undesirable cases where naive data processing indicates a high correlation of co-methylation. All such cases were removed by applying suitable filtering steps described in the data filtering section. In all three cases, the Pearson correlation of pairs of genes is r = 0.98. Because of the small variance in $\beta - values$ for C1R and LEMD3, the two genes appear highly correlated. However this kind of associations is not meaningful. All similar pairs of genes were removed from the list of results during the third stage of filtering.

Finally, we characterized the functional similarity of pairs of genes showing highly correlated methylation levels. For this, we computed the functional similarity with respect to the biological process (BPscore), molecular function (MFscore) and cellular component (CCscore) categories [182] of the Gene Ontology. The functional similarity was computed based on simRel and GOscore measures [169, 121, 183] as follows:

$$funSimAll = 1/3 \cdot [(\frac{BPscore(p,q)^2}{max_{BPscore}}) + (\frac{MFscore(p,q)^2}{max_{MFscore}}) + (\frac{CCscore(p,q)^2}{max_{CCscore}})]$$

by mapping gene names to UniProt accession identifiers with the online portal Biomyn.de.

3.2 Results and Discussion

Co-methylation analysis of more than 300 breast cancer samples from the TCGA [4] portal yielded 187 pairs of genes with Pearson correlation coefficient $|r| \geq 0.75$. These pairs were formed by 133 genes. Less than half of these pairs are located on the same chromosome. For these, we found that the level of co-methylation is weakly anti-correlated with genomic distance ($r = -0.29$). Linking co-methylation with the functional similarity of genes showed that genes with $r \geq 0.8$ tend to have similar molecular function and to be involved in the same biological process as described in the Gene Ontology. In addition to that, the found genes have high functional similarity to 22 breast cancer genes annotated in the OMIM database. Clustering of highly co-methylated genes identified six enriched KEGG pathways. Individual members of these pathways have already been linked to the progression and detection of breast cancer. Hence we have introduced co-methylation as a new tool to discover functional associations between gene pairs in breast cancer and to discover new candidate genes that should be inspected more closely in the context of the studied disease.

Pearson correlation coefficients for DNA methylation were computed for 88611328 unique pairs of genes retained after preprocessing of the raw data as described in the

[4] http://cancergenome.nih.gov/

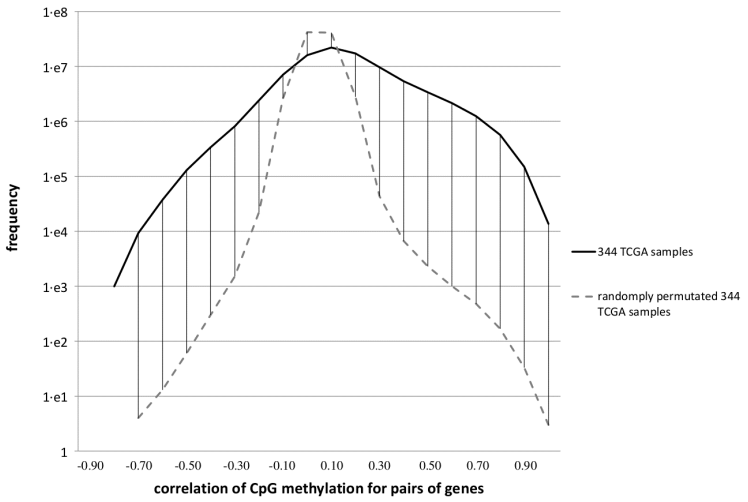


FIGURE 3.5: Distribution of different levels of co-methylation. The black curve shows the correlation computed for gene pairs in 344 breast cancer samples obtained from the TCGA portal (real data). For comparison, the gray dash curve shows the same analysis for randomly permutated samples.

Section 3.1.1. Co-methylation analysis of 183 tumor samples deposited in September 2012 (termed 'tumor group 1') yielded 98820 pairs of genes with $|r| \geq 0.75$ and 139 pairs with $|r| \geq 0.9$. For comparison, co-methylation analysis of all 344 samples (including tumor group 1 and data deposited in October 2012 termed 'tumor group 2') yielded 377547 pairs of genes with $|r| \geq 0.75$ and 13643 pairs with $|r| \geq 0.9$. The absolute frequencies of different levels of co-methylation for real data and for permutated data are shown in Fig 3.5. The co-methylation plot for randomized data on a logarithmic scale shows that $99,9\%$ of all gene pairs possess an absolute correlation of less than 0.2. In contrast, $25,41\%$ of the gene pairs show correlation higher than 0.2 for real data.

During data processing we noticed that high correlation levels may sometimes be caused by a single or a few outlying data points or may arise between genes that show very little variation in their methylation levels. Therefore we removed all gene pairs where one or both genes contained 'outlier' data points in one or more samples (see methods) and we required that all genes showed a certain variation of their CpG methylation levels. After filtering of co-methylated pairs of genes, 187 highly correlated pairs were kept (with correlation $|r| \geq 0.75$) involving 133 different genes. These gene pairs are listed in Table 3.1. Full table can be found in Appendix B.

TABLE 3.1: BE scoring of batches in BRCA adjacent normal data. The median difference counts the number of genes for which the median DNA methylation in this batch differs from its median in all other batches by a value falling into the respective intervals specified at the top.

ID	KEGG pathways	$p-value$	Genes involved in pathways	FDR
8	hsa04950: Maturity onset diabetes of the young	0.003	HNF1B, FOXA2, NEUROD1	2.622
9	hsa04640: Hematopoietic cell lineage	0.009	CD1A, CD1E, CD1D	6.229
15	hsa04730: Long-term depression	0.004	GRM5, C7ORF16, PRKG2	2.952
22	hsa04060: Cytokine-cytokine receptor interaction	0.047	EGF, TNFSF18, IL20	31.263
27	hsa04512: ECM-receptor interaction	0.005	COL5A2, COL11A1, SPP1	3.500
27	hsa04510: Focal adhesion	0.029	COL5A2, COL11A1, SPP1	17.498

TABLE 3.2: he results of disease enrichment analysis of 29 gene clusters obtained using DAVID. These clusters of genes are identical to clusters described in Table 4.4.

ID	OMIIM disease term	$p-value$	Genes involved in pathways	FDR
19	Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes	0.031	ALPL, ABO	21.382
24	Genome-wide association with bone mass and geometry in the Framingham Heart Study	0.005	CNTNAP2, KCNH8	2.345

In contrast, the data set of 344 randomly permutated TCGA samples did not contain any highly correlated gene pairs after filtering (before filtering only 91 pairs of genes passed the threshold of the correlation $|r| \geq 0.75$). The probability of identifying highly correlated methylation levels for pairs of genes in randomly shuffled data therefore equals $91/377547 = 2.41 \cdot e{-}4$. This is the $p{-}value$ for identifying highly correlated methylation levels for gene pairs in breast cancer samples.

Genomic distance

Similar to bacterial operons where neighboring genes are often expressed all at once, and in analogy to the phenomenon of genomic imprinting where a few imprinting control regions affect the allele-specific methylation in their genomic environment, one may suspect that also the methylation of neighboring genes may be more strongly correlated than that of distant genes. Here, we tested the related question whether genes showing a high correlation of their DNA methylation levels tend to be located closely to each other on chromosomes. Among the 187 pairs of genes that passed the threshold $|r| \geq 0.75$, 74 pairs of genes are located on the same chromosome. 53 out of these 74 genes are annotated in the Gene Ontology (and thus in FunSimMat) with contained BP or MF

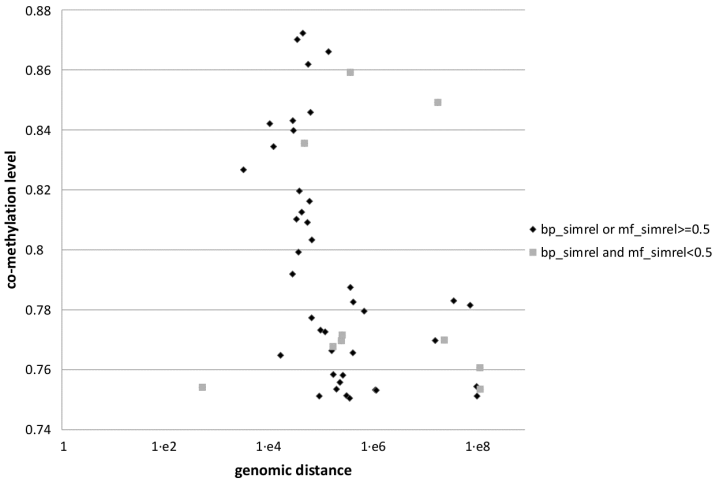


FIGURE 3.6: Distribution of different levels of co-methylation. The black curve shows the correlation computed for gene pairs in 344 breast cancer samples obtained from the TCGA portal (real data). For comparison, the gray dash curve shows the same analysis for randomly permutated samples.

scores. Figure 3.6 shows that pairs of genes on the same chromosome with strongly correlated methylation levels have a typical genomic distance between 1 e4 to 1 e6 base pairs. These values are similar to the average distance between neighboring genes of$1.4 \cdot 10^5 bp$ that is obtained when assuming that the $2.2 \cdot 10^4$human genes are evenly spread over the $3 \cdot 10^9$bases of the genome. However, the plot shows that the co-methylation level is only weakly anti-correlated with genomic distance, $r = -0.29$. For comparison, Li et al. [118] found that in human peripheral blood mononuclear cells co-methylation of CpG sites deteriorated over distance and became nearly undetectable at distances $> 1.000 bp$ [118]. We did not detect a difference between functionally similar and dissimilar gene pairs.

Functional similarity

After filtering co-methylated genes, all gene pairs were ordered according to the level of co-methylation observed. Table 3.3 shows the ten gene pairs with the strongest correlations. Interestingly, half of the cases involve two genes from the same gene family: SPRR1B and SPRR1A; FCN2 and FCN1; SPRR1B and SPRR4; REG1B and REG1P; SPRR3 and SPRR4.

TABLE 3.3: The ten strongest correlations for pairs of genes with respect to their -values, obtained after three stage filtering

First gene	Second gene	Pearson correlation
SPRR1B	SPRR1A	0.872
FCN2	FCN1	0.870
CD244	CD48	0.866
SPRR1B	SPRR4	0.862
TAS2R13	PRB4	0.859
F7	TFF1	0.856
SH3TC2	SPARCL1	0.853
ABCE1	SC4MOL	0.849
REG1B	REG1P	0.846
SPRR3	SPRR4	0.843

Next, we computed the functional similarity between the same 187 pairs of strong co-methylated genes having unique UniProt identifiers. Among these, 74 pairs involved genes located on the same chromosome and 113 involved genes on different chromosomes. Out of these, 28 and 70 pairs had complete GO annotations (Figure 3.7).

The analysis showed that, in breast cancer samples, co-methylated gene pairs on the same chromosome share a higher combined functional similarity (BP, MF, CC) than average pairs between the 133 candidate genes and the 9889 genes that are annotated in the Biomyn database out of the 14,475 genes on the Illumina27k Chip ($p{-}value_{same_{chr}} = 3.1e-4$, Welch two-tailed t-test).

In the same manner, we also analyzed the functional similarity of the 133 candidate genes to the 22 genes associated with breast cancer in OMIM, see Fig. S4. Whereas some of the candidate genes (e.g. TOX2 and GCM2) showing a large functional similarity to the 22 known OMIM genes of more than 0.8 are already being investigated with respect to breast cancer [193, 142], the 133 candidate genes as a group are less similar to the 22 OMIM genes than all 9889 genes on the Illumina27k chip with functional annotations ($p - value = 4.55e-4$, Welch one-tailed t-test). This suggests that co-methylation analysis identifies different gene players of the cellular network that are related to breast cancer on top of the 22 well-known breast cancer genes listed in OMIM.

For comparison, we also performed co-methylation analysis of the 19 OMIM breast cancer genes (out of 22) that are included in the TCGA data samples. None of them passed our strict 3 stage filtering. This is largely due to the fact that most of these genes tend to be unmethylated throughout all samples (Figure 3.8). If we leave out the last filtering condition only the 'BARD1' gene remains showing a maximum correlation value of$r = 0.626$. Interestingly, almost all OMIM breast cancer genes were methylated at low levels in the data samples that we analyzed.

Relating co-methylation to co-expression

Next, co-expression values were computed for the 187 highly co-methylated gene pairs. For this, 599 gene expression samples were downloaded from the TCGA portal and matched to DNA methylation samples by using barcodes. Altogether 336 samples were successfully matched and used for computing Pearson correlation coefficients of gene expression data. Only two pairs showed co-expression exceeding the threshold $r \geq 0.75$ (these are CD48 and SLAMF1 with $r = 0.851$; SPRR1B and SPRR2D with $r = 0.783$) and 11 pairs with co-expression $r > 0.5$. The mean Pearson correlation coefficient for the expression levels of 187 gene pairs was quite low, $r_{mean} = 0.136$. However, we found that 10 out of the 11 pairs with co-expression $r > 0.5$ are located on the

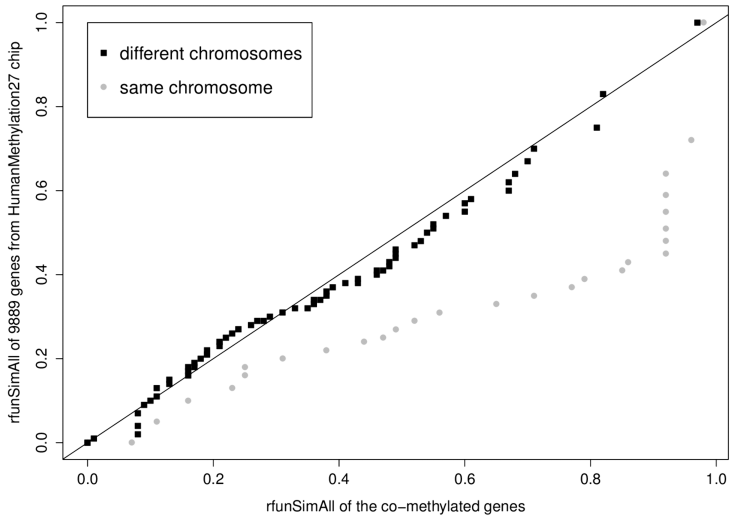


FIGURE 3.7: Q-Q plot comparing the distributions of two sets of rfunSimAll scores. Shown on the x-axis are the scores for the functional similarity among the 98 most strongly co-methylated and fully annotated pairs of genes. Here, we distinguished between gene pairs on the same chromosome and on different chromosomes. Shown on the $y-axis$ are the scores for the functional similarity of gene pairs formed between the 133 candidate genes and the 9889 annotated genes found on the HumanMethylation27 chip that could be successfully mapped to BioMyn. The semantic similarities have values between 0 (not similar at all) and 1 (totally similar) and are based on the distance of the GO terms to the lowest common ancestor in the GO hierarchy and the specificity of the lowest common ancestor.

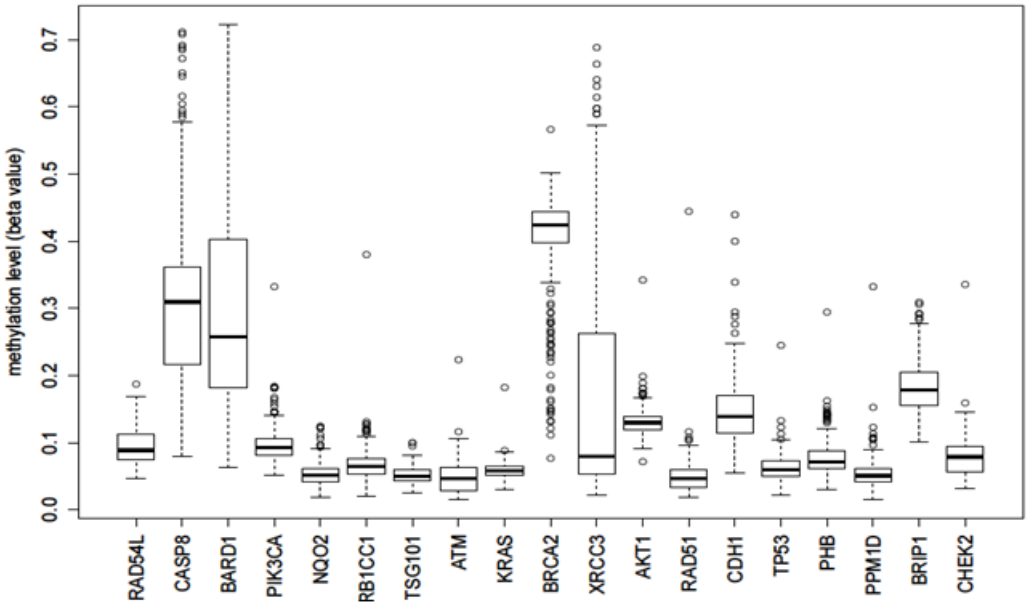


FIGURE 3.8: Methylation levels of 19 OMIM breast cancer genes found among the 344 TCGA data samples analyzed here.

same chromosome, and the 6 pairs with functional annotation have high to very high similarities (rfunSimAll is between 0.61 and 0.96).

Pathway enrichment analysis

Enriched pathways were identified by affinity propagation clustering that was performed by the 'apcluster' function in R using default parameters. Prior to clustering, three-stage filtering was applied to the methylation data of all 13313 genes. Among the 779 genes that passed the filtering steps, Affinity Propagation then identified 29 clusters of genes with similar methylation patterns. The members of each cluster were further analyzed for pathway and disease enrichment using the NIH tool DAVID [90] (Table 3.1, Table 4.4 and Appendix B Table B.1).

Among these results, we found those from pathway enrichment most interesting. Table 4.4 lists the six KEGG pathways that were significantly enriched with $p-value < 0.05$ in individual co-methylation clusters (maturity onset diabetes of the young, hematopoietic cell lineage, long-term depression, cytokine-cytokine receptor interaction, ECM-receptor interaction, and Focal adhesion). The last two pathways are related and are due to the same three genes. Interestingly, in all cases, genes belonging to these six enriched pathways have been previously shown to be associated with breast cancer.

For example, KEGG pathway 'hsa04950' relates to a form of type II diabetes termed maturity onset diabetes of the young that is caused by heterozygous mutations in at least five genes. Among these is the gene Pancreatic and duodenal homeobox 1 (PDX1) that is also known as insulin promoter factor 1.

Notably, PDX1 has been established as a marker for breast cancer by the company Roche (US patent 20070196844). Secondly, three members of the CD1 gene family, which are participants of the hematopoietic cell linage pathway, are already targets of breast cancer research [142, 70]. Moreover, CD1A has been suggested as a prognostic marker for breast cancer [110]. Thirdly, GRM5, which is a member of the long-term depression pathway, is known to be altered in breast cancer [142]. Also, gene IL20, which belongs to the cytokine-cytokine receptor interaction pathway, was shown to play a central role in the progression of breast cancer [88].

Finally, elevated expression levels of focal adhesion kinase have been associated with highly invasive human breast cancers [119, 155], focal adhesion disassembly has been linked to the potential of breast cancer metastasis [208], and ECM-receptor interaction is suggested to play an important role in carcinogenesis [106].

In conclusion, we have shown that unexpectedly strongly correlated DNA methylation levels are found in gene pairs from breast cancer patients. Importantly, correlated gene pairs show strong combined functional similarity. These findings may be helpful to annotate unknown genes and to suggest candidate genes that should be closely investigated with respect to a particular disease. We believe that our findings may also be transferable to other types of cancer, and possibly to related diseases.

We will re-visit the topic of co=methylation in the next chapter. There, we employ a novel method for correction of batch effects to avoid the problems encountered during this work where we had to filter the number of considered genes.

Chapter 4

Batch Effect detection and correction in DNA methylation data

This chapter is based on a manuscript "BEclear: batch effect detection and correction in DNA methylation data" that we submitted to the journal Bioinformatics.

When working with genome-wide high-throughput data sets e.g. for gene expression or DNA methylation, scientists encounter diverse issues connected to inconsistencies in the data. The reasons for this range from human factor over erroneous measurements, to not properly organized experiments. The well-known batch effect, a non-biological experimental variation, can be regarded as one of the vivid examples of the imprecisions that may affect the whole processing procedure of the data [158].

It can appear within one or several batches coming from experiments using high-throughput technologies making future analysis of them distorted, or even misleading. The most straightforward way to avoid batch effect is leaving out affected batches [3]. Yet, this is often not desirable since this may result in incomplete coverage of the studied issue. There exist several algorithms for detecting [25] and dealing [94] with batch effects. However, such methods typically use normalization what affects all samples in the complete dataset and might not completely remove batch effect [190]. For example, even standard normalization techniques, which are part of pipelines for transforming raw signal intensities for the DNA methylation probes into calculated $\beta - values$ mapped to the genome, might still be susceptible to batch effect (Fig 4.1).

Here, we present a novel approach for batch effect correction called BEclear. It assigns a batch effect score to every batch and exploits Latent factor models matrix approximation [103] to adjust erroneous entries. This characteristics of our algorithm is essential since it allows processing of smaller sample sizes. Moreover, we explicitly show that BEclear is able to detect not only the batches and samples that are affected by batch effect but also distinct genes responsible for that inside the samples. The correction can be applied solely to the respective genes, leaving the data for other members of the sample unchanged. We provide a detailed algorithmic framework of BEclear and report experimental results on real-world datasets demonstrating its effectiveness. In addition to this, we provide a comparison to other existing well established methods named ComBat, SVA and the recently issued Functional normalization.

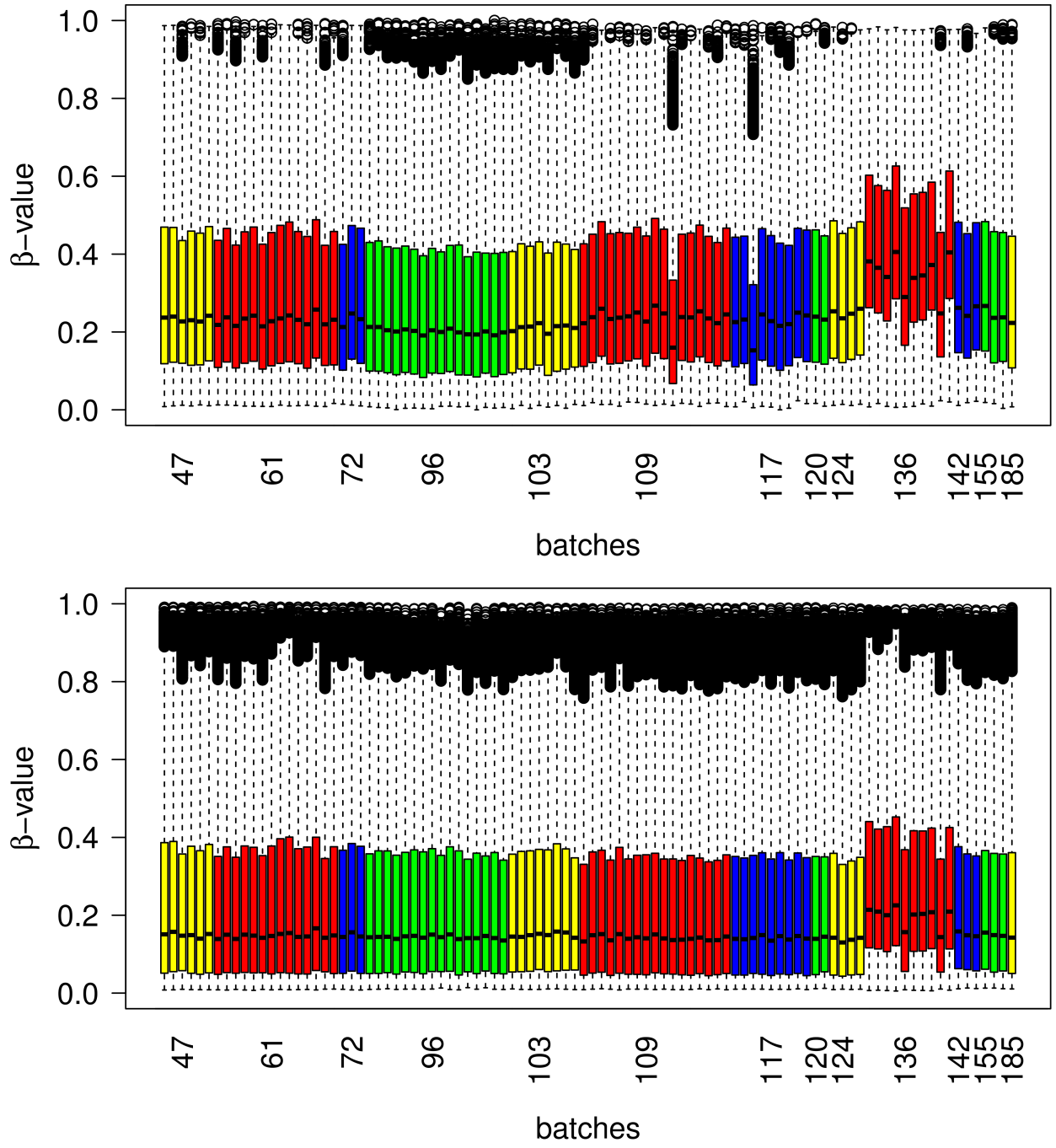


FIGURE 4.1: Per sample boxplot of adjacent normal data. A. Level 1 data - DNA methylation raw signal intensities of probes for each participant's sample. B. Level 3 data - calculated -values mapped to genome. In both cases, batch effect is clearly present in batch 136 since the distribution of $\beta-values$ in these samples significantly deviates comparing to other samples. This demonstrates the susceptibility of background correction technique from *methylumi* package to batch effect.

4.1 Materials and Methods

Tumor and Normal data

DNA methylation data for tumor and adjacent normal tissue for different tested cancer types were downloaded from The Cancer Genome Atlas (TCGA) Data Portal [1]. Level 1 (raw signal intensities of probes for each participant's sample obtained by $HumanMethylation450$ chip [17]) and level 3 (calculated $\beta - values$ mapped to the genome) array-based DNA methylation data was considered for the analysis. Our batch effect detection and correction method was established using level 3 data. However, testing was performed on both levels. We established our method on breast invasive carcinoma (BRCA) data with 745 tumor and 96 adjacent normal samples and then applied it to other cancer types.

Preprocessing of data

In a similar way as was done in [3], data from TCGA was locally stored in a MySQL database and pre-processed. Tumor and adjacent normal data were considered separately to avoid batch effects resulting from the data mixture. As a first cleaning step, all entries with missing $\beta - values$ or gene names as well as entries with indistinct gene names were removed. The next step was to keep only those probes which originate from the promoter regions of genes. For this we used annotations of the Eukaryotic Promoter Database EPDnew [50] [2] as a reference for the location of transcription start sites for every human gene. Thus, HumanMethylation450 DNA methylation probes were mapped to EPDnew data by gene name and chromosome, and only probes lying within 2000 bp up-or downstream (depending on the strand direction) were kept for further analysis. Some genes were still represented by multiple probes in a single sample file. For those genes we assigned the mean $\beta - value$ of all its respective entries. This resulted in 11154 gene $\beta - value$ pairs in tumor matched data and 11213 in adjacent normal.

4.1.1 Batch effect detection and correction method BEclear

Visual inspection of the data

At the beginning of this procedure, it is essential to establish a standard for the grouping of samples into batches. Here we used the batch identifier from the TCGA data portal to assign every single sample to its respective batch. In order to find out whether the data is affected by batch effects, several standard visualization approaches were applied separately to tumor and adjacent normal samples, namely box plots, density plots, heat map together with hierarchical clustering, and principal component analysis (PCA). These well established methods are very good at representing data globally to get a general impression on batch effects in a single batch or even distinct sample. But they do not reveal whether single genes within a group of samples belonging to the same batch are affected or not.

Detection of single batch effected genes (BE-genes)

Genes within a batch that are likely affected by batch effects were discovered by applying statistical analysis based on the comparison of batch medians. Since it is unclear whether DNA methylation data is distributed e.g. according to a normal distribution, we

[1] http://cancergenome.nih.gov/
[2] http://epd.vital-it.ch/

used the nonparametric Kolmogorov-Smirnov test. Iteratively going through all batches, the distribution of every gene in one batch was compared to the distribution of the same gene in all other batches. The p-value returned by KS-test was then corrected by FDR [16] All genes with significant $p-value$ below 0.01 were considered for further analysis.

Next, to consider only biologically relevant differences in methylation levels, the medians of the $\beta-values$ for the previously identified genes were computed in exactly the same manner as when applying the KS-test. Thereby, we identified the median difference the absolute difference between the median of all $\beta-values$ within a batch for a specific gene and the respective median of the same gene within all batches. Those genes from our list with median difference above 5%of $\beta-value$ distribution ($mdif \geq 0.05$) that also passed the KS-test were considered as batch effected genes in a specific batch. Importantly, every batch has its own list of BE-genes.

Batch effect scoring (BE-score) and correction

After single BE-genes were found, the decision about batch effect correction can be made. The scoring for the batch effect in a dataset was computed for every batch according to the formula:

$$BEscore = \frac{\sum\limits_{i \in mdif_{cut}} (N_{BEgenes_i} \cdot W_i)}{N}$$

where N is the total number of genes in a current batch, is the category of median differences, $mdif_{cut}$ is the number of BE-genes belonging to the $i-th$ $mdif$ category and w_i is the weight for the respective $mdif$ category. Weights were assigned in the following way:

- if $mdif < 0.05$, then $weight = 0$;
- if $0.05 \leq mdif < 0.1$, then $weight = 1$;
- whenever $mdif$ takes values in the interval [0.1; 1], with step size 0.1 w_i is increased by two.

This formula considers not only the number of BE-genes in the batch, but also the deviation of the medians of BE-genes in one batch compared to all other batches. Thus, the higher the BE-score for a batch is, the more this batch was affected.

If at least one of the batches has a high BE-score, then all BE-gene entries in respective batches should be adjusted. This was done by removing them and then performing matrix completion using Latent Factor Models (LFM) based on matrix factorization [103, 27]. The main advantage of this method is the ability to incorporate both gene and sample preferences by taking into account the values of neighbor entries when predicting a missing value.

Assuming that the dataset is represented by the $m \times n$ matrix D, given rank r, LFM iteratively constructs an $m \times r$ matrix L and an $r \times n$ matrix R such that matrix multiplication $[LR_{ij}]$ approximately equals to $[D_{ij}]$ for every unaffected entry A.

The Gradient descent optimization algorithm was used to minimize the global loss i.e. the difference between $[LR]$ and D. When it converges, non batch effected entries were preserved in the completed matrix $D_{comp} = LR$ from the original data matrix D, so that the algorithm affects only the matrix entries for BE-genes in some of the batches. In case if some of the predicted entries lie below 0 or above 1, they were assigned 0 and 1, respectively.

4.1.2 Method validation

Different values were tested for $mdif$, $p-value$ and $p-value$ adjustment methods used during the detection of single batch effected genes in adjacent normal BRCA data. In case when $mdif$ was too strict (larger than 0.1), only few genes (from 103 to 1465 BE-genes) were detected as BE-genes. After removing them, batch effect was still visually observed. In contrast if $mdif$ was 0.01 then it detected more than 82% of all genes as BE-genes. Different thresholds for the p-value did not affect the results so strongly. With a $p-value = 0.05$ Beclear identified 5990 BE-genes and for $p-value = 0.001$ 5032 BE-genes. All three $p-value$ adjustment methods (FDR, Hommel and Bonferroni) yielded approximately similar numbers of BE-genes which is around 5500 genes.

The proposed method for matrix completion was assessed from the perspectives of overall accuracy and prediction time when applied to DNA methylation data. For testing purposes again the BRCA adjacent normal dataset was used. As a measure of accuracy, we computed the average absolute deviation between known and predicted entries of the matrix. Due to the fact that BEclear found 5.8% entries to be affected, testing was performed on 6% of additionally randomly selected entries.

Generally, the time needed to perform LFM prediction grows exponentially with the size of the data. For the BRCA dataset studied here (11213 genes in 96 samples), this task could be infeasible without separating the initial matrix into blocks of data and running LFM independently for every block. This approach gives an additional advantage since it allows usage of parallel computing using multi-core processors, what leads to significant savings in computation time.

We also analyzed how the size of the block of data to which LFM was applied affected the prediction accuracy. This parameter was varied from 10 to 250 in increments of 5. In all cases LFM yielded a similar accuracy which is in the range of 0.02 (Figure 4.2).

Note that in case when the size of the block of data is too large, this significantly affects the computation time without bringing an improvement in accuracy. From another point of view, a very small block size might not incorporate gene preference since there might be some large batch with batch effect. And the block could contain some inner part of that batch.

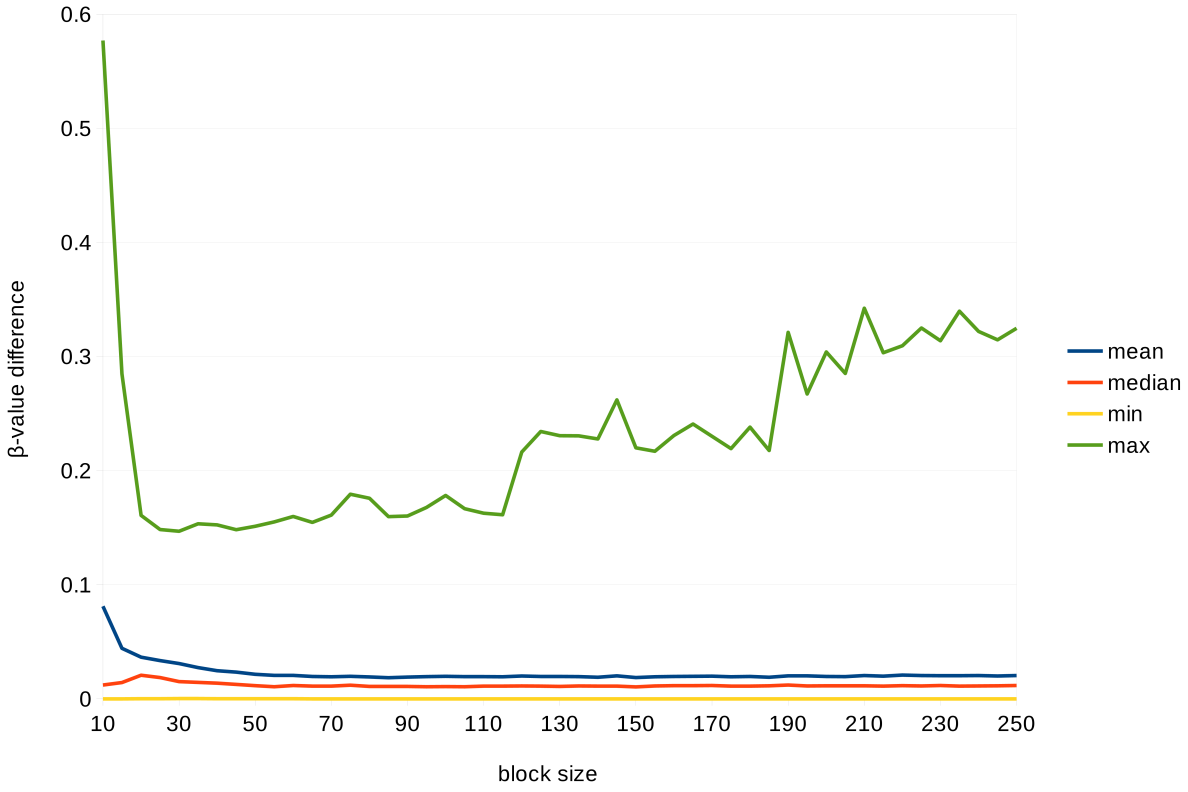


FIGURE 4.2: Latent Factor Model accuracy assessment. Investigating the impact of the block size on the overall accuracy of LFM matrix completion. Four parameters were computed: mean, median, minimal and maximum of the difference between actual and predicted $\beta - value$ entries. The size of the block of the data, to which LFM was applied, changes from 10 to 250. With the increase of the block size, the chance to get few high $\beta - value$ differences also grows. However, LFM shows good prediction accuracy since median of difference remains in the range of 0.01 whereas the mean stays around 0.02.

4.2 Results and Discussion

4.2.1 Box plots and further visual analysis of BRCA data

The BEclear method is currently tailored towards level 3 data for genome-wide DNA methylation data. We will illustrate its performance using data for breast cancer samples from the TCGA portal [3].

Box plots representing the distribution of $\beta - values$ (proportion of methylated CpG nucleotides ranging from 0 to 1) for all genes were generated both on a per sample and a per batch basis (Figure 4.3).

These plots illustrate clearly that the distribution of $\beta - values$ for genes in batch 136 from the BRCA samples is noticeably increased compared to the other batches.

[3] http://cancergenome.nih.gov/

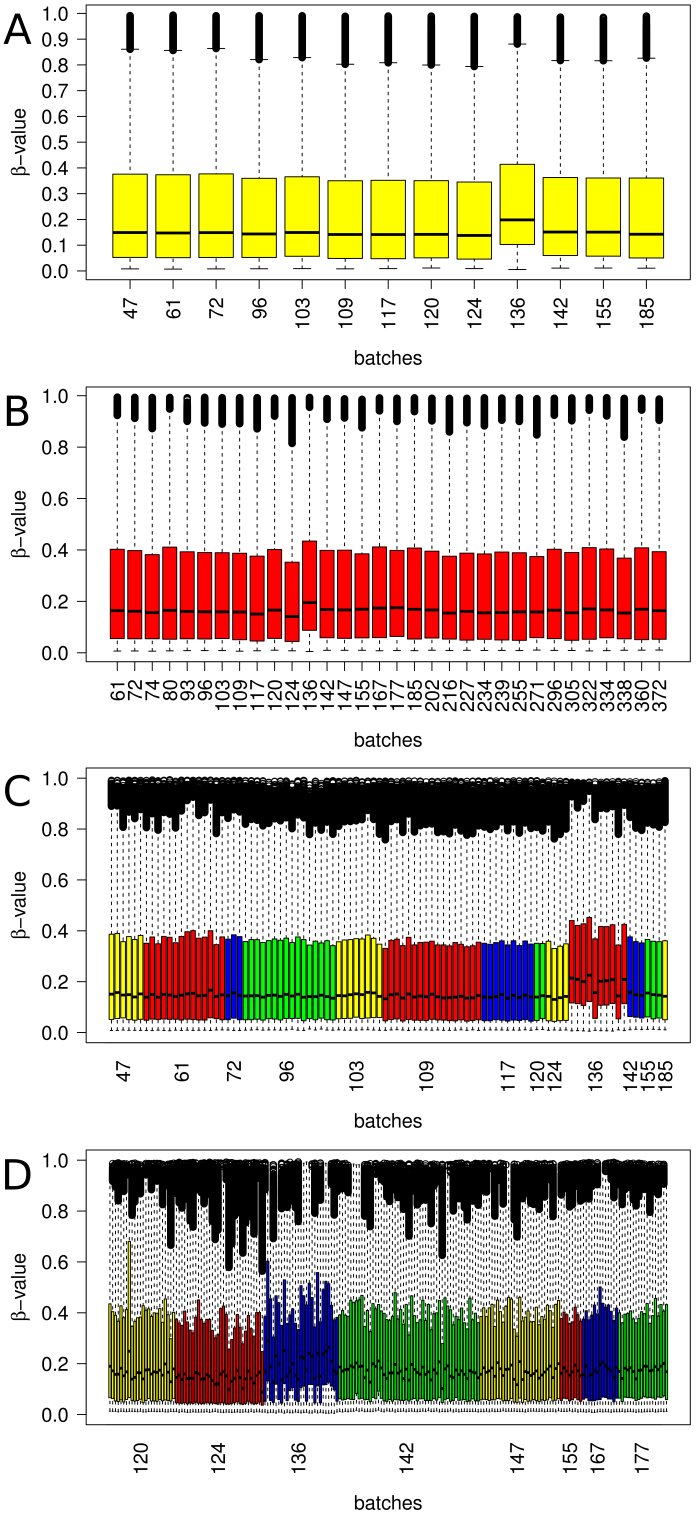


FIGURE 4.3: Box plots of breast cancer samples. A. Adjacent normal samples per batch level (13 batches). B. Tumor samples, per batch level (32 batches). C. Adjacent normal samples, per sample level (96 samples). D. Subset of tumor samples for batch 136 and surrounding batches, per sample level. All these plots illustrate clearly that batch 136 is affected by batch effect in both tumor and adjacent normal samples.

Particularly, in the adjacent normal data the first quartile, median and third quartile are increased by more than 0.05 compared to all other corresponding values from the other batches. The per sample plot (Fig 4.3 C) shows that the difference in batch 136 is not due to only one sample but exists in all except two samples within this batch.

Also the tumor data (Fig. 4.3 B) of batch 136 show a general increase of $\beta - values$. However, the difference is not as large as in the adjacent normal data, as seen in the per sample plot, where only 15 out of 27 samples behave differently compared to other batches. This may reflect that tumor data has inherently more variation in the distribution of the $\beta - values$ than normal data.

This batch effect in adjacent normal data was also well apparent in the PCA, heatmap and density plots (Fig 4.4.).

Clearly, most of the batch 136 samples tend to cluster together (Fig 4.4 A, B) and the density of this batch is less sharp and shifted compared to other batches (Figure 4.4C).

4.2.2 Batch effect detection and correction results in BRCA data

As just explained, both tumor and adjacent normal samples in the BRCA data from the TCGA portal contain a batch effect in batch 136.

TABLE 4.1: BE scoring of batches in BRCA adjacent normal data. The median difference counts the number of genes for which the median DNA methylation in this batch differs from its median in all other batches by a value falling into the respective intervals specified at the top.

batch ID	[0.05;0.1)	[0.1;0.2)	[0.2;0.3)	[0.3;0.4)	[0.4;0.5)	BE-score
47	91	32	4	0	0	0.015
61	274	63	8	0	0	0.039
72	6	5	1	0	0	0.002
96	33	2	0	0	0	0.003
103	13	0	0	0	0	0.001
109	143	5	0	0	0	0.014
117	93	3	0	0	0	0.009
120	3	0	0	0	0	0
124	14	1	0	0	0	0.001
136	3992	1159	104	9	1	0.605
142	10	0	1	0	0	0.001
155	8	0	1	0	0	0.001
185	0	0	0	0	0	0

This result observed by visual inspection was also confirmed by the new BEclear method introduced in this manuscript. Table 4.1 lists the number of BE-genes in every batch separated by the median difference mdif. For example, the distribution of the SPINK2 gene in batch 136 (Figure 4.5) is statistically significantly different from the distribution in all other batches, as confirmed by Kolmogorov Smirnov test ($p - value = 9.41 \cdot e - 6$). The difference between the median $\beta - value$ for this gene in batch 136 and the median in all other batches is in the range of $[0.4; 0.5)$. Table 1 shows that all except one batch contain some BE-genes. However their number is typically relatively small as well as the median deviation, what leads to a small BE-score. It is immediately noticeable that this dataset clearly is affected by a strong batch effect in

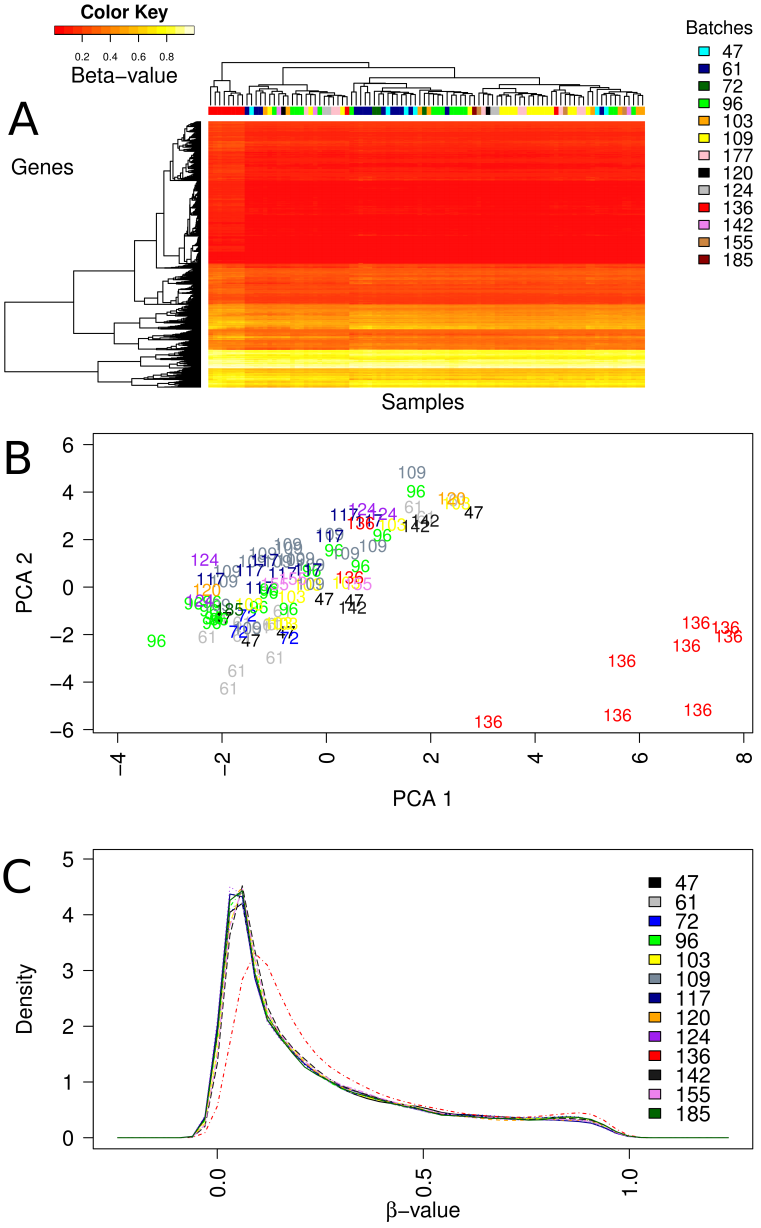


FIGURE 4.4: Visual inspection of batch effect in adjacent normal Breast Invasive Carcinoma data. A. The heatmap demonstrates that all but two samples from batch 136 form a cluster that splits off from the other samples at the top of the hierarchy. B. Plotting the first two Principle Components and projecting samples on them clearly distinguishes batch 136 samples from the rest. C. The density plot of every batch shows the difference between the distribution of $\beta - values$ of all genes in batch 136 compared to all other batches

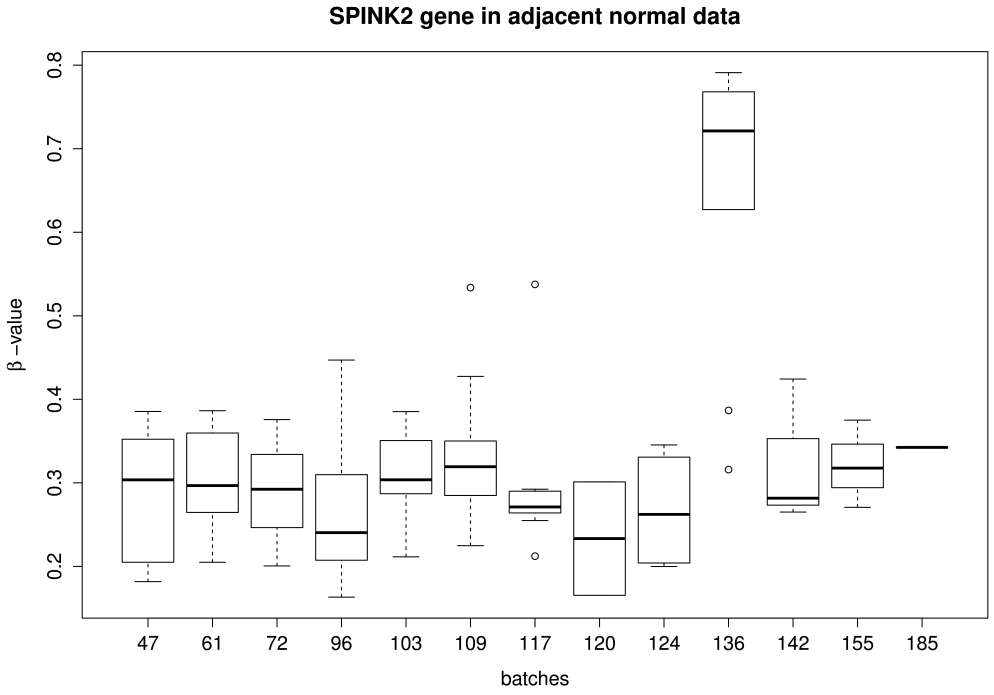


FIGURE 4.5: Per batch boxplot of the $\beta-values$ for gene $SPINK2$ in adjacent normal data. For this gene we identified the largest median difference of 0.428.

batch 136 since approximately 47% of all genes in this batch differ from the same genes in the other batches with respect to $\beta-value$ by more than 0.05. This median difference can reach up to 0.5 resulting in a BE-score = 0.6. A similar picture is observed in BRCA tumor data. There the batch effect in batch 136 is not as drastic as in adjacent normal data but still has a BE-score = 0.19 (Table 4.2).

The high BE-score for batch 136 in breast cancer data suggests the necessity of applying a batch effect correction method to this dataset such as the one presented in this chapter. BEclear adjusted the methylation values of 6079 genes in 12 batches in adjacent normal data and 3587 in 31 batches in tumor data. The results are shown in Figure 4.6. In the per batch boxplot (Fig. 4.6 A) batch 136 does not stand out explicitly anymore what is also confirmed by the per sample boxplot (Fig. 4.6 B). However, it is also apparent that a certain variation between samples remains since BEclear adjusted only BE-genes. Even though the tumor dataset had a smaller batch effect than adjacent normal samples, it was successfully adjusted and now the bar corresponding to batch 136 is in a similar range compared to other batches (Fig. 4.6 C). Additionally, Fig. 4.6 D and E confirm the positive effect of BEclear on normal data. The corrected 136-th batch data is now positioned next to all other batches.

TABLE 4.2: BE scoring of batches in BRCA tumor data

batch ID	[0.05;0.1)	[0.1;0.2)	[0.2;0.3)	[0.3;0.4)	BE-score
109	37	2	0	0	0.0037
117	25	3	0	0	0.0028
120	12	4	0	0	0.0018
124	166	13	1	0	0.0176
136	1661	199	10	1	0.1887
142	19	3	0	0	0.0022
147	14	2	0	0	0.0016
155	6	2	0	0	0.0009
167	13	1	0	0	0.0013
177	69	11	0	0	0.0082
185	1	2	0	0	0.0004
202	4	1	0	0	0.0005
216	1	0	0	0	0.0001
227	13	2	0	0	0.0015
234	3	1	0	0	0.0004
239	17	2	0	0	0.0019
255	12	2	0	0	0.0014
271	31	7	0	0	0.004
296	3	2	0	0	0.0006
305	12	3	0	0	0.0016
61	165	24	0	0	0.0191
322	102	31	5	2	0.0176
334	300	93	16	7	0.0531
338	16	2	0	0	0.0018
72	45	5	0	0	0.0049
74	79	10	0	0	0.0089
80	189	26	2	0	0.0223
93	0	0	0	0	0
96	18	0	0	0	0.0016
103	9	1	0	0	0.001
360	13	3	0	0	0.0017
372	24	6	1	0	0.0036

Batch effect scoring of other tumor types

To show the general applicability of the BEclear method to DNA methylation data, 7 further well represented cancer types from the TCGA portal were assessed toward batch effect. Here only level 3 data was considered for the analysis. Beside breast invasive carcinoma, which was discussed above, BEclear only identified a minor batch effect in tumor samples of Kidney renal clear cell carcinoma, KIRC (Table 4.3). Interestingly, a similar findings was recently reported by Fortin et al. [64]. As in the case of BRCA data, KIRC has relatively many batches but batch 32 is represented by only 2 samples (Figure 4.7 A). Even though this batch doesn't contain many BE-genes, the median difference mdif of those genes is very large (Figure 4.7 B.), yielding a BE-score of 0.185.

This analysis gives the clue to the question how large should the BE-score be in order to perform batch effect correction of the data. Hence, BE-score greater than 0.1 is a strong signal toward the presence of batch effect in the dataset. Another conclusion suggests that the more batches exist and the smaller they are, the higher is the chance of finding a batch effect there.

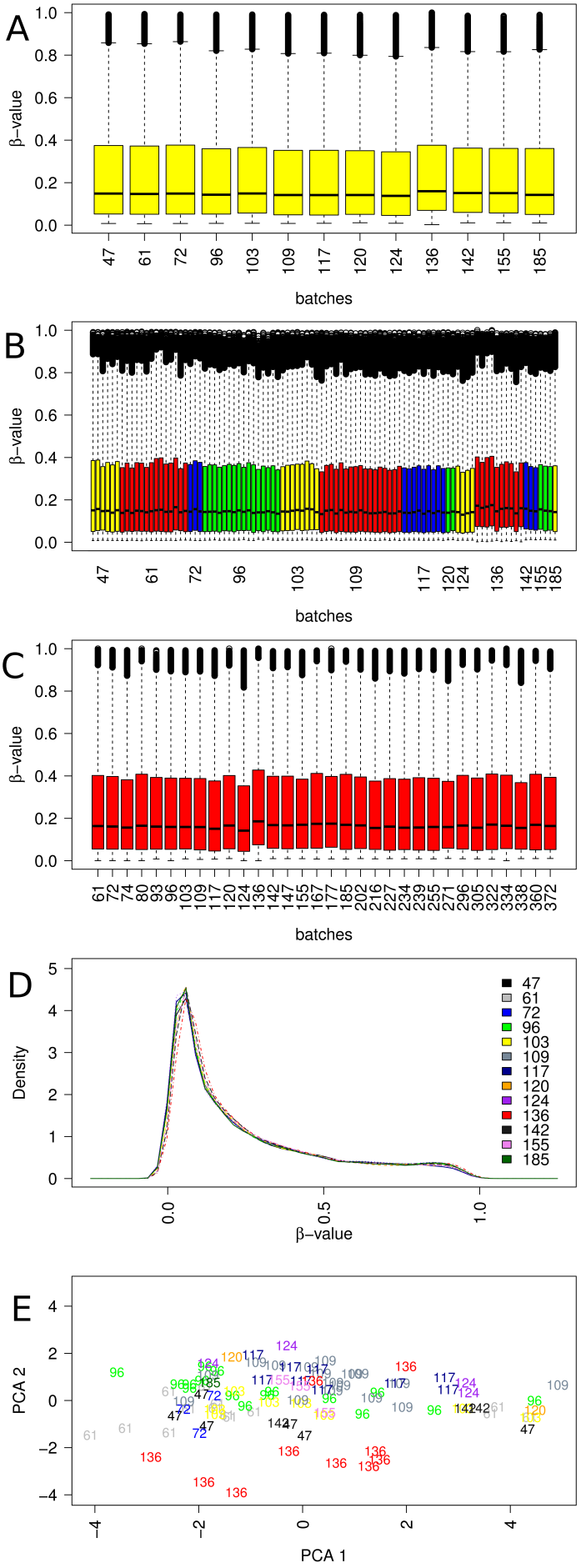


FIGURE 4.6: Results of batch effect correction of breast cancer data using BEclear. A. Per batch and B. Per sample boxplot of adjacent normal data. C. Per batch boxplot of tumor data. D. Density plot and E. PCA plot of adjacent normal data.

TABLE 4.3: BE scoring of 7 different cancer types from the TCGA portal. Cancer types and batches which were confirmed to have batch effect are marked in bold. This table contains the description of cancer types, batch identifiers obtained from TCGA portal and batch effect score. Only those batches with BE-score over 0.01 are listed here since, generally, every batch has some extremely small non-zero BE-score. This is due to some variation in a few genes and only in rare cases the BE-score for a batch is exactly zero. All the batches belonging to Lung squamous cell carcinoma have BE score in a range of (0; 0.01) because not more than 97 genes in a single batch behave differently compared to other batches.

batch ID	BE score
Breast invasive carcinoma BRCA,adjacent normal data, represented by 13 batches	
47	0.015
61	0.038
109	0.013
136	0.605
BRCA, tumor data, represented by 32 batches	
61	0.019
80	0.022
124	0.017
136	0.188
322	0.017
334	0.053
Uterine Corpus Endometrial Carcinoma UCEC, adjacent normal data, represented by 12 batches	
104	0.027
UCEC, tumor data, represented by 23 batches	
49	0.021
92	0.036
156	0.015
186	0.017
Thyroid carcinoma THCA, adjacent normal data, represented by 12 batches	
115	0.014
THCA, tumor data, represented by 17 batches	
115	0.016
Kidney renal clear cell carcinoma KIRC, adjacent normal data, represented by 5 batches	
82	0.042
KIRC, tumor data, represented by 12 batches	
32	0.185
387	0.037

batch ID	BE score
Head and Neck squamous cell carcinoma HNSC, adjacent normal data, represented by 4 batches	
83	0.0108
107	0.0106
151	0.0143
HNSC, tumor data, represented by 18 batches	
107	0.013
145	0.015
260	0.0107
265	0.032
403	0.019
Lung adenocarcinoma LUAD [4], adjacent normal data, represented by 4 batches	
37	0.015
LUAD, tumor data, represented by 18 batches	
37	0.027
52	0.016
84	0.02

4.2.3 Comparison against existing BE correction methods

Next, we compared BEclear against several existing well established methods for batch effect correction. ComBat [94] is used for batch effect adjustment of microarray expression data and is a part of the Surrogate Variable Analysis package [115] in *R* [5]. ComBat uses an empirical Bayes framework based on a location (mean)/ scale (variance) model and assumes that batch effects can be corrected so that all batches have similar values of means and variances in all batches.

Since DNA methylation data generally does not follow a normal distribution, the decision was made towards the nonparametric version of ComBat to correct BRCA data. Before running the batch effect adjustment, batches 185 and 93 were excluded from normal and tumor data, respectively, because ComBat is not able to handle batches with just one sample.

Both adjacent normal and tumor (Figure 4.8 data were corrected separately by ComBat. The tool was obviously able to remove the observed batch effect in batch 136 by equalizing upper quartiles, medians and lower quartiles for every box in normal data. In contrast to the adjacent normal data, the variation between the range of the boxes in tumor data is mostly maintained compared to the original data whereas the formerly outstanding batch 136 is obviously corrected and boxes are shifted to a similar level compared to the other batches. Inspection of the number of BE-genes remaining after BE correction showed that both ComBat and BEclear were able to remove batch effect and had a similar performance (Figure 4.9).

Nevertheless, we noticed that ComBat has important drawbacks and unwanted effects which are not present in BEclear. Previously, we mentioned that ComBat cannot handle

[5] http://www.r-project.org/

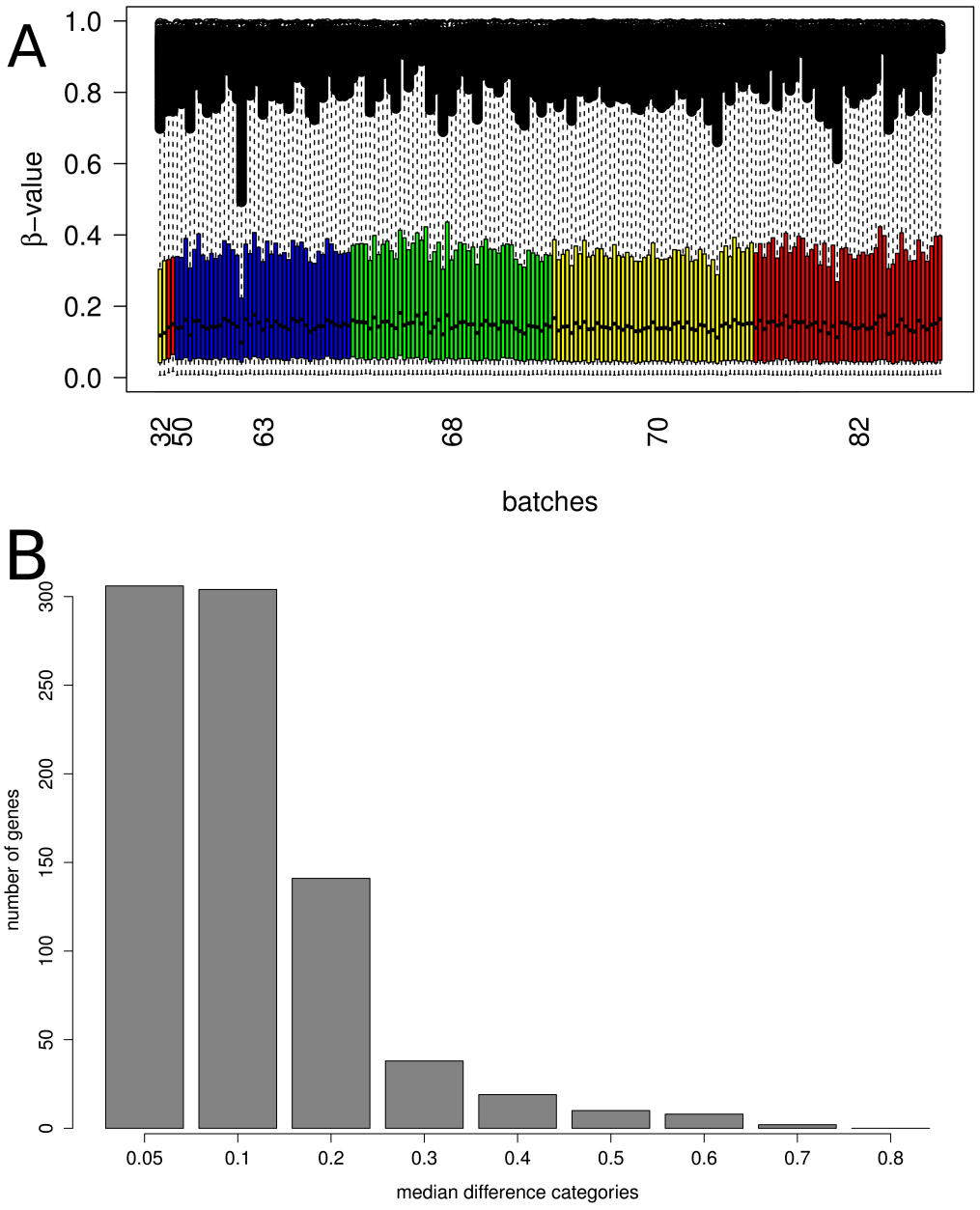


FIGURE 4.7: DNA methylation data for kidney renal clear cell carcinoma tumor samples, KIRC. A. Per sample boxplot. Batch 32, which is represented by two samples, has a batch effect score equal to 0.185 signaling that his data should be corrected. B. The number of genes belonging to different categories of median differences ($mdif$) between genes in the current batch and the same gene in all other batches (as described in the section 'Batch effect scoring').

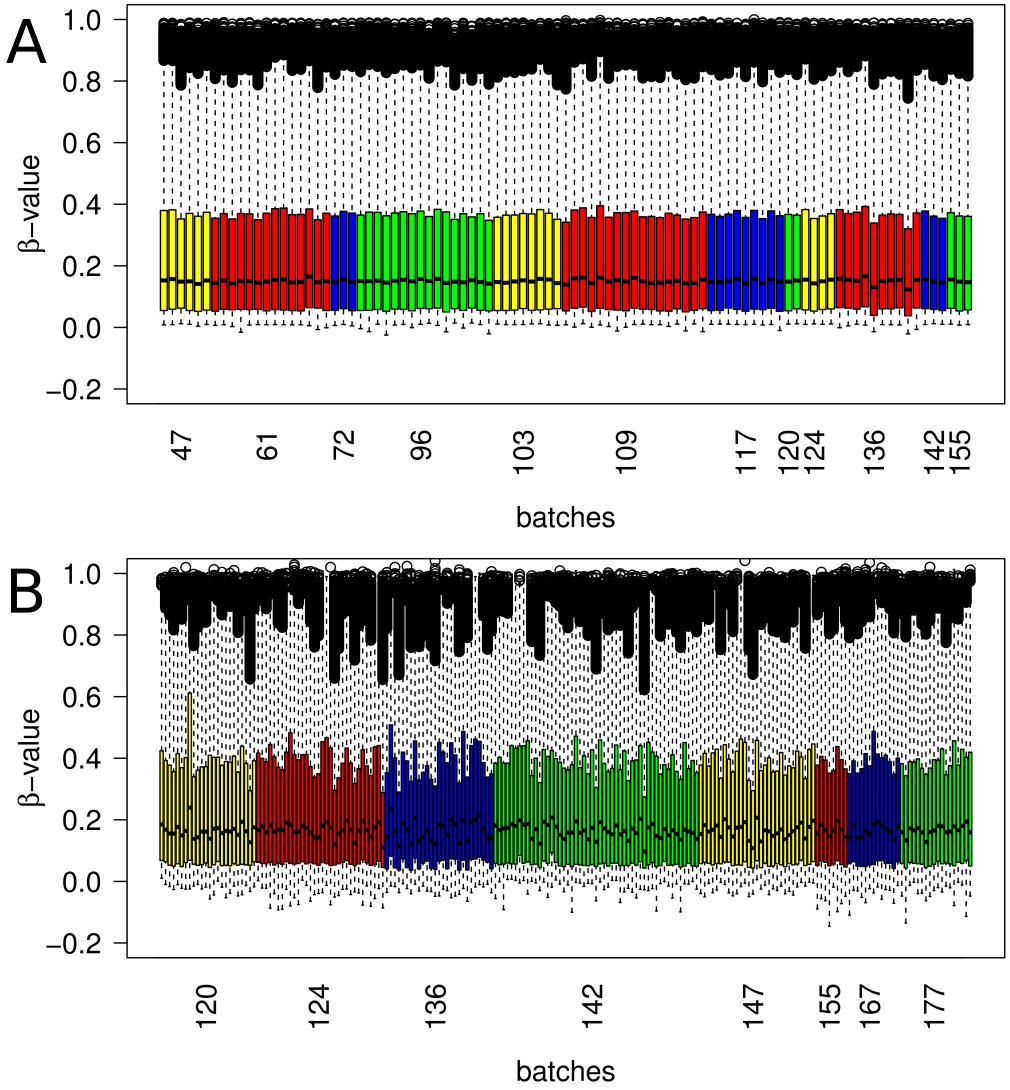


FIGURE 4.8: Removing batch effect in BRCA data using ComBat. The previously observed batch effect in batch 136 was corrected both in A. adjacent normal and B. tumor data.

batches that only contain a single sample and requires that the data should follow a normal distribution. More importantly, ComBat adjusts all entries in the dataset even though not all of them are affected by batch effect.

Especially in the tumor data, which inherently contains more variation, this eventually could smoothen the data too much and might diminish biological variation to some extent. In contrast, BEclear leaves all unaffected parts of the data as is and only replaces batch effected genes in some specific batches by the predicted entry based on the gene and samples preference. Another artifact, which should be pointed out, is that ComBat produces many values above 1 and below 0, what should not be the case, because the distribution of $\beta value$ should stay in the range of [0;1] (Figure 4.10).

For tumor data after BE correction, ComBat produced 261 values above 1 and 6529 below 0. In contrast, BEclear yielded only 32 for every case. But more important is that the undesired entries generated by BEclear do not exceed the interval [0;1] by more

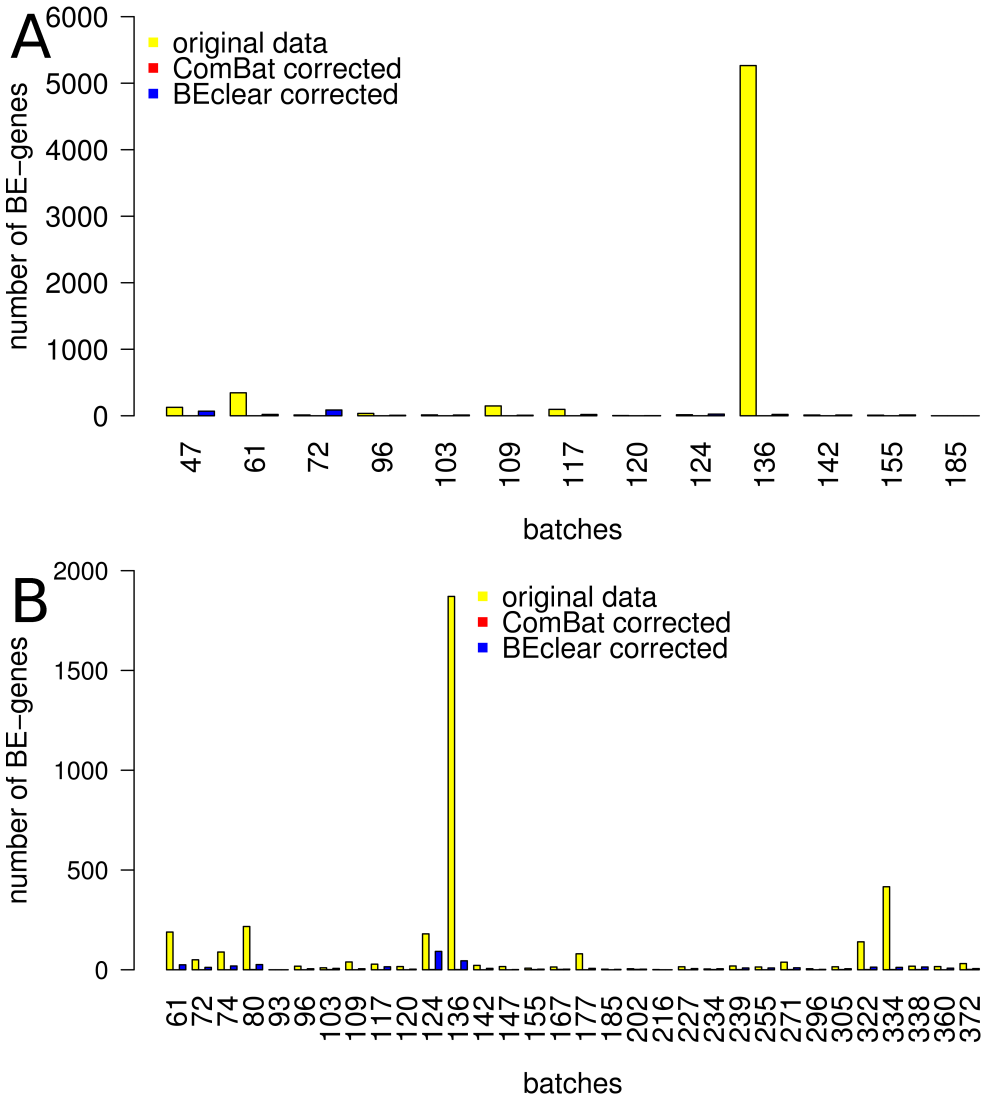


FIGURE 4.9: Comparison of original data, data adjusted by ComBat and data adjusted by the proposed BE correction method. Shown are the counts for the number of batch effected genes. A. BRCA adjacent normal and B. BRCA tumor data were used. As a measure, the number of BE-genes in every single batch was used.

than 0.06 whereas for ComBat those values can be beyond 0.15. In the case of adjacent normal data only few such cases were observed - ComBat didn't give any value above 1, whereas BEclear had only 3 entries. However, ComBat returned 37 entries below 0 whereas BEclear had no violating entries in this category. This could be explained by the fact that ComBat was designed to handle batch effects in gene expression data whereby the value range is not restricted to stay between 0 and 1. In cases, where most of the genes are unmethylated, it will shift the data too much towards 0 resulting in many entries lying below 0 (Figure 4.10 B.). Such problems arise with BEclear much less often. We finally eliminated this problem by cutting values at 0 and 1.

Another method against which we benchmarked BEclear was Surrogate Variable Analysis (SVA). Since it still uses ComBat as a correction tool, it has the same drawbacks as discussed above. When applying SVA to level 3 adjacent normal BRCA data, we

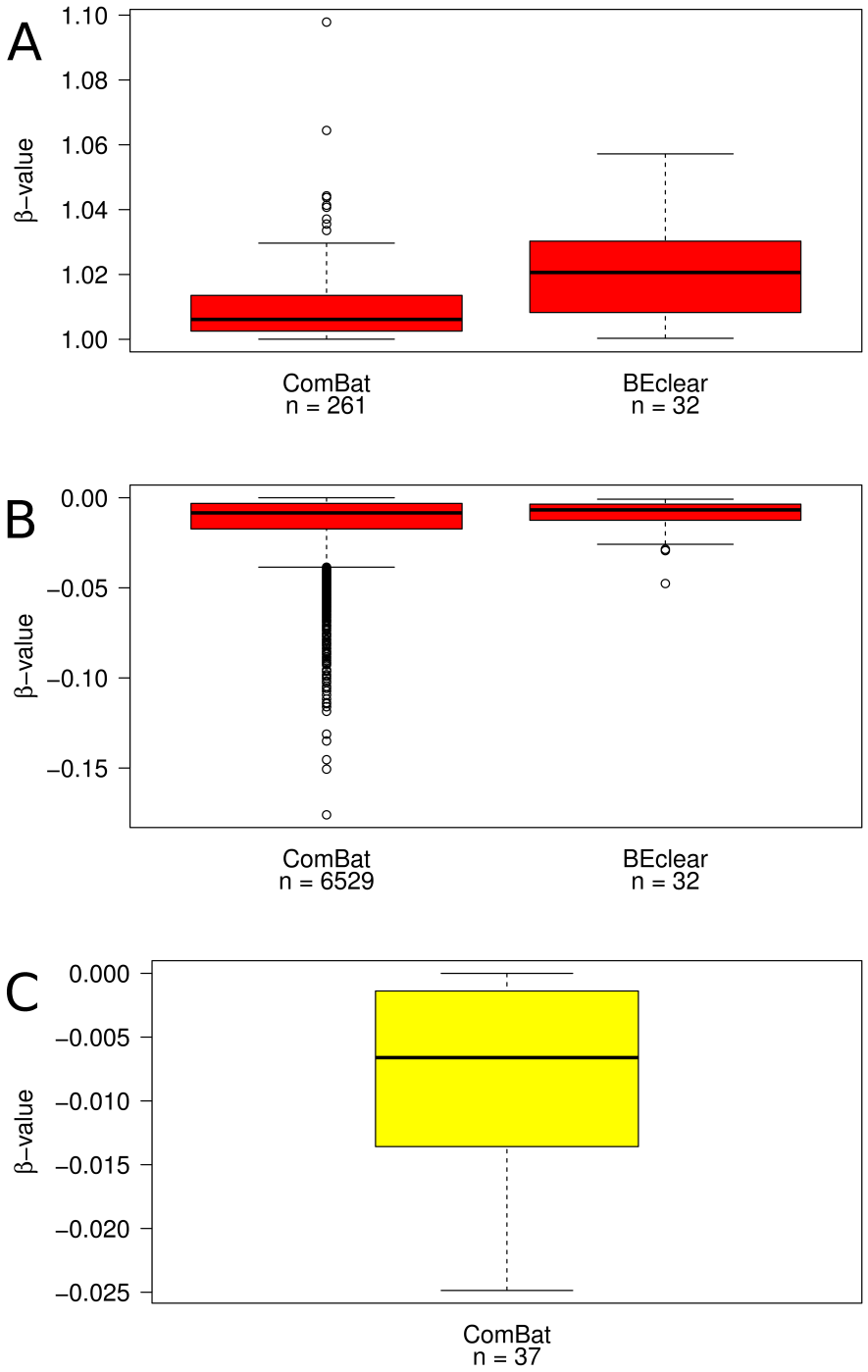


FIGURE 4.10: Comparison of ComBat and BEclear with respect to the number of wrongly predicted entries below 0 or above 1. A. Boxplot of entries which are above one and come from the batch effect adjusted tumor data. Both ComBat and BEclear were applied. B. The same as in A showing the number of values below 0. C. Boxplot of values below 0 of adjacent normal data after correction by ComBat.

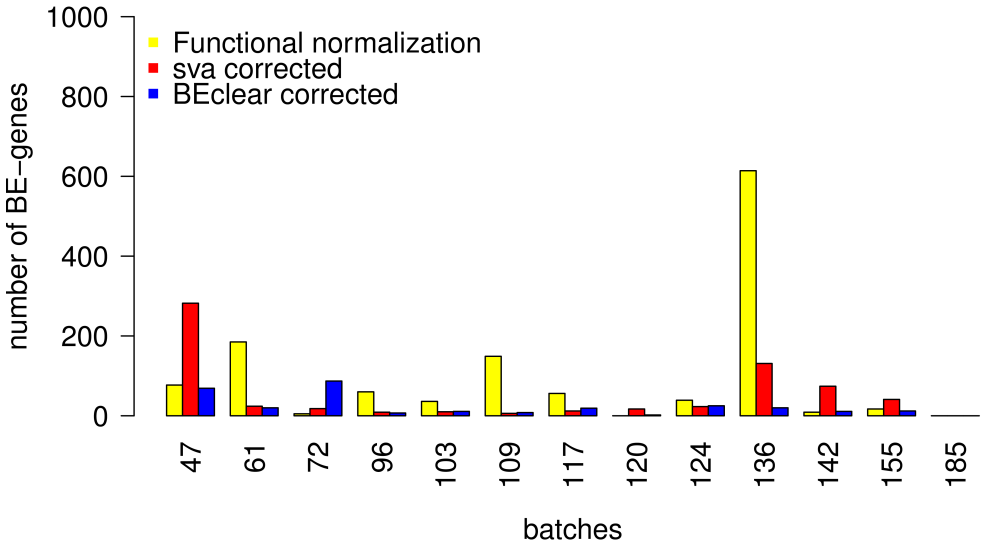


FIGURE 4.11: Comparison of BEclear, SVA and Functional normalization (*minfi* package) with respect to the number of BE-genes still remaining after the correction. According to this criteria, BEclear was able to outperform the two other methods.

noticed that it was able to remove batch effect to a large extent still preserving variation in the data, in distinction to ComBat. However, computing the number of BE-genes left after batch effect adjustment, SVA was outperformed by BEclear (Figure 4.11).

Finally, BEclear was compared to Functional normalization, which was introduced recently and was designed specifically for 450k methylation array. This method is part of the minfi package [10] and can work only with level 1 data. Thus following the pipeline introduced in this package we reached level 3 data and then preprocessed this data as described in the respective section of the current paper. This method was indeed able to remove batch effect from the first prospective (Figure 4.12 A), but then the density of batch 136, the most affected group of samples, still differs from the density of other batches ($p - value = 4.03 \cdot e - 4$, Figure 4.12 B).

When counting the number of BE-genes remaining after batch effect correction, functional normalization didn't reach the performance of BEclear having 1128 BE-genes (755 out of which identified in the batch 136, the most affect batch) in contrast to 223 BE-genes (20 from batch 136), respectively (Figure 4.11). Another important fact we observed is that almost half (1353 out of 3804) of all human housekeeping genes (HKG) [56] are affected by batch effect what leads to an increase of the methylation level in the most affected batch 136. Since generally HKG promoter regions should be unmethylated, we studied their behavior in the adjacent normal BRCA data before batch effect correction and after applying BEclear or functional normalization (Figure 4.13).

Especially focusing on those 1353 batch affected HKG it is clearly seen that batch 136 is still shifted slightly upwards after functional normalization what is not the case for BEclear where all the bars have approximately equal first, third quartiles and median.

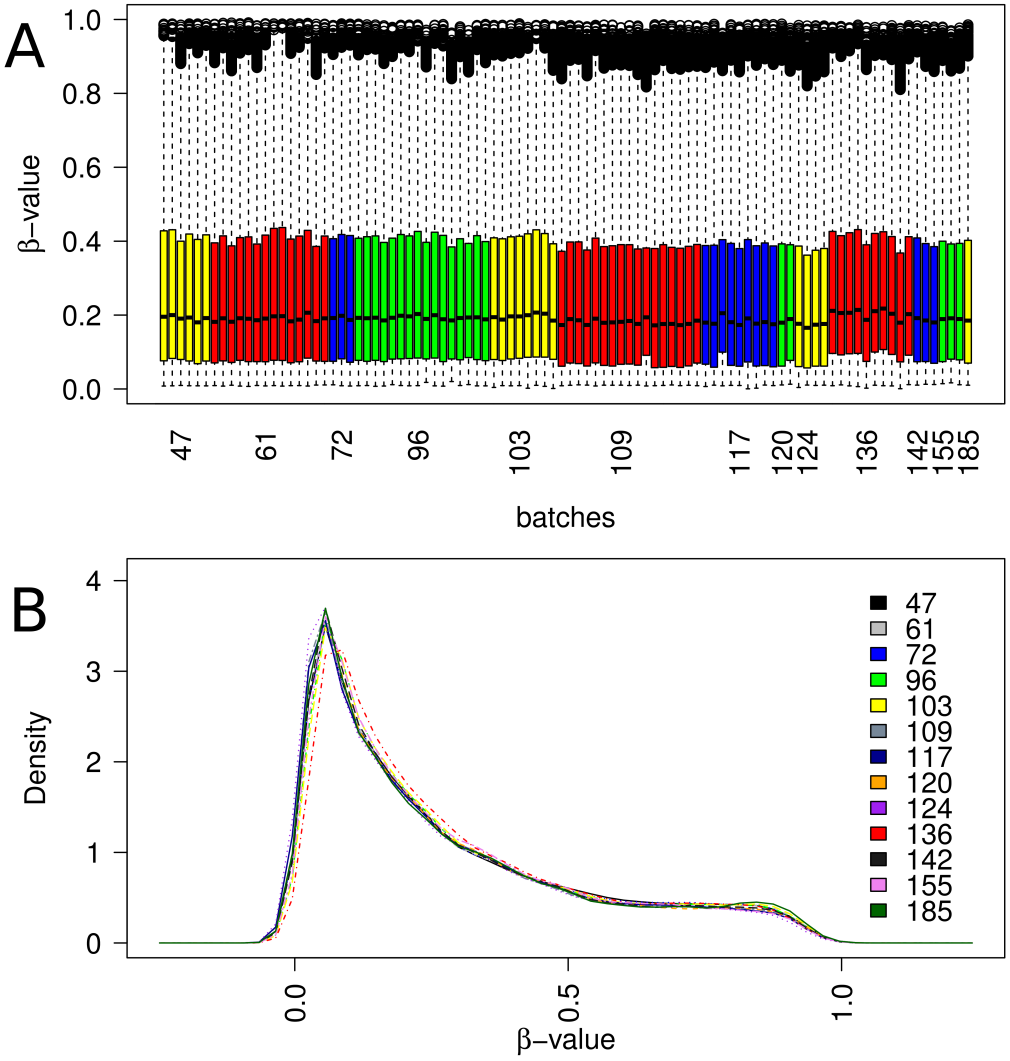


FIGURE 4.12: Results of batch effect adjustment of breast cancer adjacent normal data using Functional normalization. A. Per sample boxplot B. Density plot. Functional normalization was able to adjust batch effect well and Fig 11 A looks very similar to what was obtained after BEclear correction (Figure 4.6B).

4.2.4 Co-methylation and differential methylation

Co-methylation analysis was performed in the same manner as in our previous work [3] on BRCA data before and after applying BEclear, in order to investigate the impact of batch effect on the amount of artifacts. Since the data is already preprocessed and contains promoter region methylation, then only pairwise Pearson correlation and 3 step filtering are left. Co-methylation analysis was performed for three sorts of data: adjacent normal, tumor and combination of tumor and adjacent normal data. The number of tumor samples significantly exceeds the number of adjacent normal samples, hence only samples coming from the same participants were considered for the combined dataset. They were matched by TCGA barcodes resulting in 190 samples all together. One of the filtering steps suggests excluding batch effected genes from the analysis, however this step was avoided because co-methylation was applied on both kinds of datasets with and

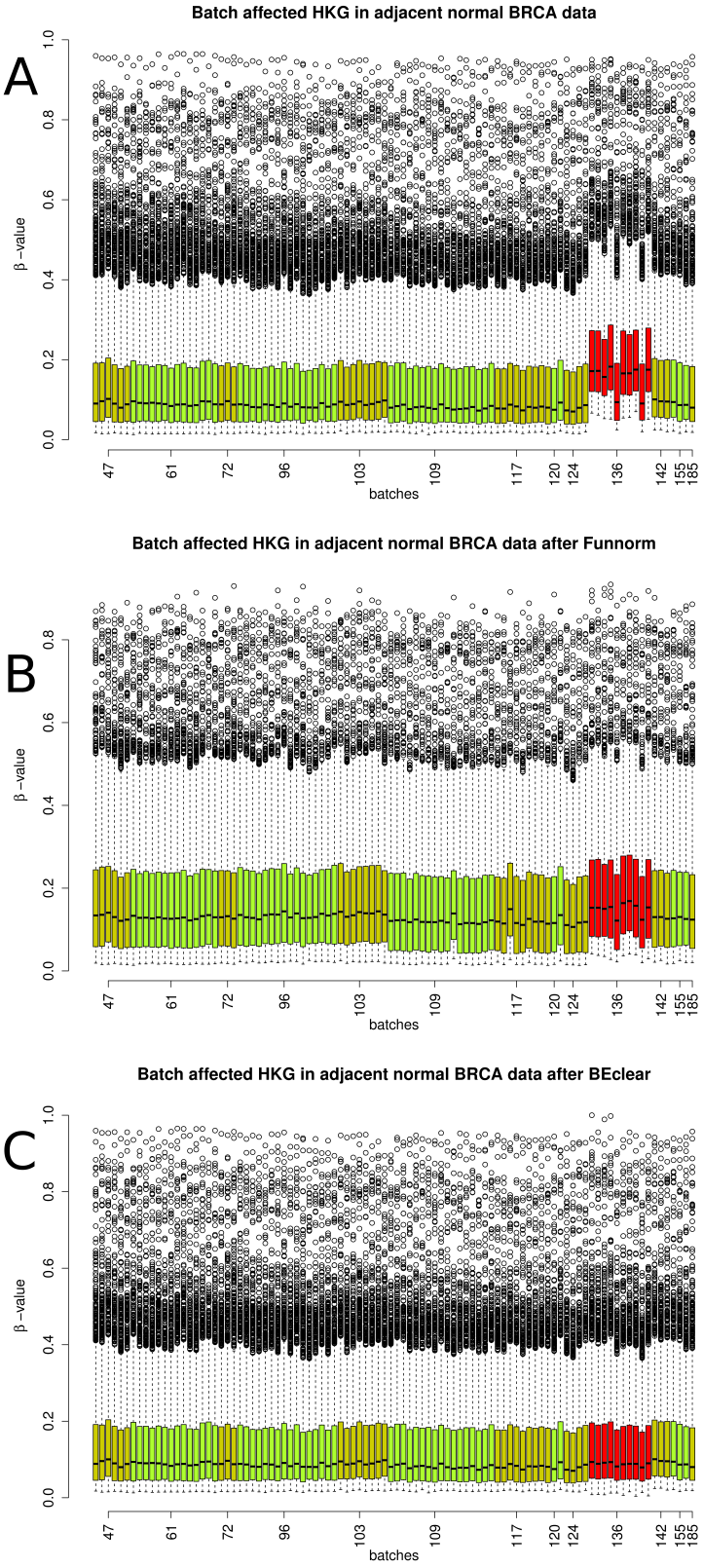


FIGURE 4.13: Boxplots of 1353 batch affected housekeeping genes in adjacent normal breast invasive carcinoma data A. before any batch effect adjustment B. after functional normalization C. after BEclear batch effect correction. The most affected batch is marked in red.

without batch effect. Table 4.4 contains the number of pairs of genes with correlation higher than 0.75 or lower than −0.75 for different datasets. Clearly, batch effects are responsible not only for generating false associations between genes with respect to their methylation levels in different samples, but also for losing a large portion of expected relationships. This behavior doesn't depend on the data type and can be observed in tumor, adjacent normal and combined samples.

TABLE 4.4: Statistics representing co-methylation results. This table contains the number of highly co-methylated pairs of genes for three different types of data after batch effect adjustment and before.

	Tumor samples	Adjacent normal samples	Combined samples
Total number of pairs of genes in dataset before BE correction	115	8206	9592
Total number of pairs of genes in dataset after BE correction	112	4517	10616
Number of common pairs of genes in BE-corrected and uncorrected datasets	112	4228	8893
Number of pairs of genes present in corrected dataset and absent in uncorrected	0	289	1723
Number of pairs of genes present in uncorrected dataset and absent in corrected	3	3978	699

Finally, differential methylation analysis between tumor and normal samples was carried out applying Kolmogorov Smirnov test (KS-test) [19, 206, 53, 202] and Significance analysis of microarrays SAM [195, 116] for 190 combined adjacent normal and tumor samples. The KS-test returned the list of genes whose distribution in normal samples differs from the distribution in tumor samples with $p - value$ below 0.01. To verify this list, SAM was applied independently and only genes returned by both methods were considered for further analysis. In this way two lists of differentially methylated genes were generated one for data without batch effect correction and another for data after applying BEclear. These lists contain 6147 and 6672 genes, respectively, resulting in 616 genes which were present in the data after batch effect adjusting only. Inspecting these genes one can find genes which are known to play an important role during cancer development or even have been associated with breast cancer before: NRG4, TUBB, LPL, BRD2, MYB, RAP2C, SIRT7, MAZ, HRAS, TXN, PPM1D, TP53I3, PARK7, TP63 [81, 97, 107, 4, 164, 76, 11, 203, 204, 83, 120, 24, 30, 143, 98].

We have shown that batch effect may cause missing differentially methylated genes and the neglect of batch effect adjustment may generally be a barrier towards discovering important associations in cancer studies.

Chapter 5

Matched-cohort DNA microarray divercity analysis of methicillin sensitive and methicillin resistant *S.aureus* isolates from hospital admission patients

This chapter is based on the publication entitled "Matched-cohort DNA microarray diversity analysis of methicillin sensitive and methicillin resistant *Staphylococcus aureus* isolates from hospital admission patients" by Ruffing et al. in PLoS One journal (2012) [174].

Staphylococcus aureus is notorious as a major human pathogen causing invasive disease such as deep abscess formation, endocarditis, osteomyelitis, and sepsis [126]. The health care system of the world is challenged by the presence of methicillin-resistant *S. aureus* (MRSA). Diverse strains of it have been characterized based on genetic profiles for healthcare associated (haMRSA), community associated (caMRSA) [72] and also for livestock associated infections (laMRSA) [99, 100].

A lot of efforts have been put to to associate *S. aureus* gene profiling [93, 124, 141] of clonal lineages with either ecological success [178] or clinical disease [80]. However, genetic traits responsible for rendering a given *S. aureus* clone are still to be outlined. the main point to overcome MRSA in medical buildings must be a drastic reduction of its transmission. This control needs a precise information regarding the source and spread of nosocomial pathogens.

Yet, this information is limited with regard to prevalent healthcare associated MRSA strains, as the typically clonal albeit regionally divergent phylogenetic traits of prevalent isolates [74] often preclude in-depth transmission pattern analyses. Moreover, the lack of routinely accessible information on the virulence gene equipment prevents any attempt for differentiated therapeutic or infection control approach as a function of pathogen equipment [174].

Hygiene management is highly dependent on genotyping of *S. aureus*. Standardised and fast methods are needed for effective and fast evaluation of closely related epidemic

strains. In this chapter data obtained from a commercially available DNA microarray (IdentiBAC) is compared with standard *spa*-typing for *S. aureus* genotyping. A subgroup of 46 MRSA and matched 46 MSSA nasal isolates were collected within a state-wide admission prevalence screening in he Saarland University Medical Center.

Author contributions The data collected for these samples was provided to us by our collaborators Prof. Dr. med. Mathias Herrmann, Prof. Dr. med. Lutz von Müller and Ulla Ruffing. My task was the bioinformatics analysis of this data except for the splits graph analysis that was performed by Ulla Ruffing.

Normally *spa*-typing along with Microarray (MA) are capable to distinguish genetically diverse MSSA groups. Yet, due to the predominance of CC5/t003 samples in the MRSA group, a more detailed subtyping is needed for more complex genetic profiles analysis.

The genetic repertoire of the MRSA group is characterized by more virulence genes as compared to the MSSA group. The standard evaluation of MA results by the original software into *CCs, agr-, SCCmec-* and capsule-types was substituted by implementation of multivariate subtyping of closely related CC5 isolates using three different bioinformatic methods (splits graph performed by Ulla Ruffing, cluster dendrogram, and principal component analysis). Every of these approaches was applicable for standardized and highly discriminative subtyping with high concordance. We assumed that the identified *S. aureus* subtypes with characteristic virulence gene profiles are presumably associated also with virulence and pathogenicity in vivo. However, this is to be studied in more detail in the future.

5.1 Preliminaries

Simplicity and standard processing with the help of easy software tools have made genomic analysis of the variable X-region of the *S. aureus* protein A gene *(spa)* [102, 188] by single locus sequencing (*spa*-typing) very popular. However, the discriminatory power of this analysis has limited abilities in an epidemiological setting. One can use it as a frontline tool for *S. aureus* typing exclusively with additional discriminatory markers as e.g. *SCCmec* typing, lineage-specific genes or specific gene polymorphisms [188, 43]. Usage of Multilocus sequence typing (MLST) together with DNA macrorestriction results in even smaller numbers of distinguishable genotypes.

Multiple-locus variable-number tandem-repeat analysis (MLVA) [66, 85, 184] is able to provide distinction even within similar genotypes. At the same time it includes numerous steps of sequencing which involve expensive consumables and equipment.

Complete genome analysis with the help of next generation sequencing is being successfully applied for outbreak analysis [104] and in the nearest future will still remain an application for specialized laboratories. When applied to a specific cluster (e.g. the *t003* type) analysis of single nucleotide polymorphism (SNP) is able to further differentiate with a high discriminatory power, yet, in general each SNP probe is unique and restricted to respective clonal complexes [150].

One promising perspective which provides a reasonable compromise between easy applications, cost and adequate time limits is a commercial diagnostic DNA-based MA panel (Alere IdentiBACH StaphyType Microarray [*IdentiBAC MA*]). Is is specifically aimed for *S. aureus* genotyping [139]. This approach consists of the comprehensive analysis of the *S. aureus* genome by hybridization to 334 different genetic probes. This

offers highly reproducible simultaneous analysis of 174 genes distributed over the complete *S. aureus* genome [134, 135, 51]. Genes which are being analysed can be combined into lineage specific *S. aureus* groups: resistance and virulence genes. Consequently, *agr-*, *capsule-* and *SCCmec* typing as well as a highly accurate discrimination of *S. aureus* lineages is implemented [173, 132].

Crude IdentiBAC MA results are available and MA analysis has been applied for a broad collection of MRSA isolates [134], reporting 34 MRSA lineages and more than 100 different strains in human as well as veterinary isolates.

Here, IdentiBAC MA data is used for the first time in a subgroup of MRSA and matched MSSA isolates collected within a large, state-wide admission prevalence screening in the State of Saarland. All isolates of MSSA colonized patients matched based on gender, age and previous hospital admissions were included as a control group of patients with similar predisposition and exposition to healthcare associated infections [174]. MA analysis is augmented with spa-typing for independent lineage attribution, and results are subjected to advanced bioinformatic analysis. The following questions are to be answered:

- What is the clonal lineage distribution of MSSA and MRSA isolates across a time and region-restricted hospital admission screening?
- Can one observe any a difference in the accessory gene equipment of MRSA and MSSA hospital admission- associated isolates?
- Are there differences between bioinformatics models in respect to phylogenetic lineage delineation?
- Does bioinformatics analysis help to further differentiate between predominant clones indistinguishable by *spa*-typing and clonal complex (CC) attribution?

The next section 5.2 gives further details on data and methods used.

5.2 Materials and Methods

Patients and Clinical Isolates

Clinical isolates were collected in a 4 weeks interval during routine hospital entry screening from patients with nasal *S. aureus* colonization admitted to the Saarland University Medical Center. 6 MRSA isolates and 46 matched isolates of the MSSA colonized control group were included. Matched controls were selected according to gender, age ($< 70 vs. \geqslant 70 years$), previous hospitalizations in general and in the last 6 months (Table 5.1). Criteria were selected to match patients with a similar risk exposure for community and healthcare associated *S. aureus* contacts. The study was approved by the ethic commission of Saarland (registration No. 127/10).

TABLE 5.1: Risk factors of MRSA and matched MSSA control group isolates.

Risk factors	MRSA, n (%)	MSSA, n (%)	$p-value$
Male	18 (39.13%)	18 (39.13%)	$*$
Female	28 (60.87%)	28 (60.87%)	$*$
$< 70 years$	24 (52.17%)	24 (52.17%)	$*$
$\geqslant 70$ years	22 (47.83%)	22 (47.83%)	$*$
Hospitalisations < 6 months	21 (45.65%)	21 (45.65%)	$*$
Inter-hospital transfer	5 (10.64%)	1 (2.17%)	ns
Previous MRSA colonization	3(6.52%)	1(2.17%)	ns
MRSA contacts	8(17.39%)	4(8.70%)	ns
Long-term care	11(23.91%)	2(4.26%)	0.014
Retirement home	3(6.52%)	0(0.00%)	ns
Diabetes mellitus	9(19.57%)	8(17.39%)	ns
Antibiotic therapy	21(45.65%)	8(17.39%)	0.007
Dialysis	3(6.52%)	0(0.00%)	ns
Medical devices	8(17.39%)	0(0.00%)	0.006
Skin lesions	6(13.04%)	2(4.26%)	ns

Spa-typing

DNA of clinical isolates was prepared by boiling ($95 \circ C$ for 10 minutes) followed by amplification of the polymorphic X region of the protein A gene *(spa)* using standard primers *spa*-1113f (5' TAA AGA CGA TCC TTC GGT GAG C 39) and *spa*-1514r (5' CAG CAG TAG TGC CGT TTG CTT 39). Before sequencing (ITseq, Kaiserslautern, Germany) the PCR product was digested by Exo- SAP ITH (Affymetrix, Cleveland, United States) at $37 \circ C$ (15 minutes), and the reaction was terminated at $80 \circ C$ (15 minutes). Sequences were assigned into *spa*-types using the Ridom StaphType software version 2.1.1 and BURP algorithm (Ridom GmbH, Münster, Germany), as described previously [79].

DNA Microarray-based Genotyping

DNA extraction and hybridization to the IdentiBAC MA (Alere Technologies GmbH, Jena, Germany) was performed as described in the manufacturers instructions [14, 137] . In brief, genomic DNA was purified using the cell lysis components of the assay in combination with DNeasy blood and Tissue kit (Qiagen, Hilden, Germany). The test principal is based on a linear multiplex primer elongation using one primer for every single target and DNA labeling by incorporation of biotin-16-dUTP. Following DNA hybridization, microarray probes were washed, then horseradish- peroxidase-streptavidin precipitation reaction was performed resulting in visible grey spots in case of a positive reaction. Spot signals were recorded, and automatically analyzed using the designated ArrayMate reader and the corresponding software (Iconoclust, Alere Technologies) [14]. As result, the MA readings of 334 target sequences corresponding to 174 distinct genes were classified into species markers, genes encoding virulence factors, microbial surface components recognizing adhesive matrix mol- ecules (MSCRAMMS), antimicrobial resistance genes or SCC*mec*-, capsule- and *agr*- typing markers. As part of the IdentiBAC MA results in conjunction with the Iconoclust analysis, array profiles are attributed to a specific clonal complex (CC) and sequence type (ST) based on a proprietary algorithm provided by the manufac- turer. Similarly, SCC*mec* types are attributed as a result of array signals obtained.

Splits Graph Construction

A network tree was constructed by splits graph analysis (SplitsTree 4.11.3 software, *www.splitstree.org*) which was automatically linked to *spa*-typing results based on the computed export cost/distance matrix using the BURP algorithm of the Ridom StaphType software. The microarray results were imported directly into SplitsTree software 4.11.3 [91], and analyzed on default settings (characters transformation, uncorrected P; distance transformation, Neighbour-Net; and variance, ordinary least squares).

Cluster Dendrogram Construction

Phylogenetic-like analysis of microarray hybridization pattern profiles was performed using R (version 2.13.1, $http://www.r-project.org/$) in conjunction with Bioconductor packages [68]. First, the data were preprocessed by removing all gene IDs containing ambiguous results. Afterwards, genes can only be present (1') or absent (0') in a particular sample. Next, the Euclidean distance matrix was computed to measure the similarity of gene hybridization profiles in different samples using the dist function in the software package "Stats R, version 2.13.1). Finally, a cluster dendrogram was constructed employing the hierarchical agglomerative clustering method and using by the hclust function in "Stats" that is based on Wards method [162, 161].

Principal Component Analysis

As a multivariate analysis, principal component analysis (PCA) was carried out for *S. aureus* MA results to reduce the dimensionality of the MA data, and to identify groups of correlated variables. PCA characterizes the degree of variability (variance) observed among the detected genes. It combines the data for individual genes into so-called principal components (PCs) that are ordered according to the magnitude of variance observed in the data. Projecting the full data set onto the first few PC vectors showing the largest variance then allows a powerful reduction of data without loosing much information. The same preprocessed data was used as in the clustering analysis. PCs were computed by the R function prcomp in package "stats" with default parameters and the options retx = TRUE, center = TRUE and scale = FALSE). By definition, the first principal component is the particular linear combination of gene hybridization profiles that contains the largest variation in the data. The second PC is the linear combination of the hybridization profiles that explains the largest variation after removing the first PC and so on. Here, only the first two PCs were considered for the present analysis.

Statistics

Statistical evaluation was done by non-parametric tests using Fishers exact test.

5.3 Results

Patients and Clinical Isolates

Patient characteristics were matched between the MRSA and the MSSA group for the selection criteria (sex, age, previous hospitalizations) whereas significant differences were found between groups for history of long-term care, previous antibiotic therapy, dialysis and the presence of medical devices (Table 5.1).

***spa*-typing**

The 46 MRSA isolates were assigned to 13 different *spa*-types (Table 6.1). The predominant MRSA spa-type was the epidemic strain *t*003, Rhine-Hesse $(29, 63\%)$. A higher diversity was uncovered among the 46 MSSA-isolates classified into 33 different spa-types

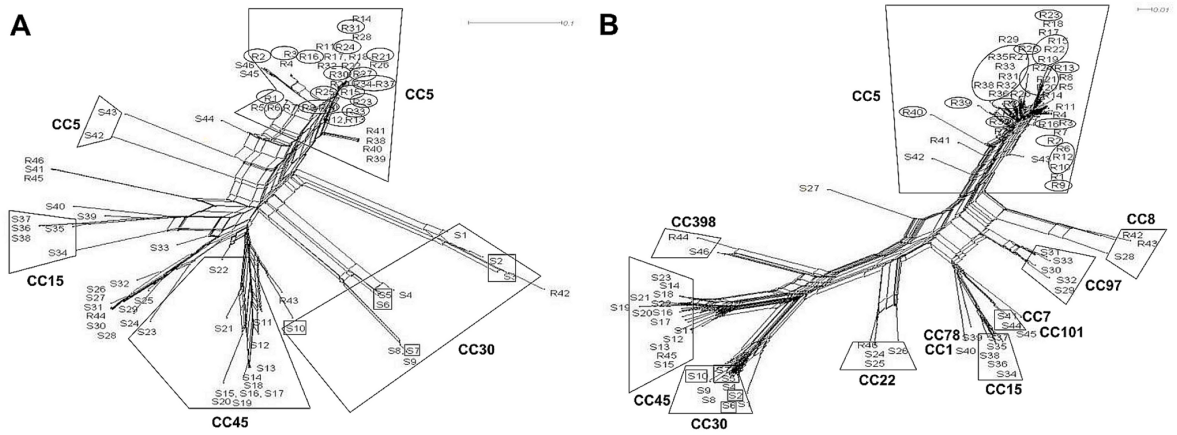


FIGURE 5.1: Diversity analysis of all MSSA (S1S46) and MRSA (R1R46) isolates by splits graph. (A) Splits graph constructed based on cost distance matrix produced by Ridom StaphType and (B) on default settings of the IdentiBAC microarray hybridization profiles of 334 genes and alleles. Clonal complexes (CC) as well as the most abundant *spa*-types t003 (circles) and t012 (quadrates) were highlighted.
doi:10.1371/journal.pone.0052487.g001

with the most common MSSA *spa*-types being $t012$ $(6, 13\%)$ and $t015$ $(5, 10.9\%)$. For MSSA, *spa*-typing allowed for good discrimination of patient isolates which was shown here by splits graph analysis; however, the majority of MRSA isolates clustered into CC5/$t003$ which hampered sub-classification by *spa*-typing (Figure 5.1A).

TABLE 5.2: Differences of spa-types and clonal complexes in MSSA and MRSA isolates.

Clonal complex	*Spa* -type	MRSA, n (%)	MSSA, n (%)
CC1	t8864	1 (2.17%)	0 (0%)
CC5	t003	0 (0%)	29 (63.04%)
	t504	0 (0%)	4 (8.70%)
	t010	0 (0%)	2 (4.35%)
	t002	1 (2.17%)	1 (2.17%)
	t045	0 (0%)	1 (2.17%)
	t481	0 (0%)	1 (2.17%)
	t493	1 (2.17%)	0 (0%)
	t887	0 (0%)	1 (2.17%)
	t1079	0 (0%)	1 (2.17%)
	t3195	0 (0%)	1 (2.17%)
CC7	t091	2 (4.35%)	0 (0%)
CC8	t008	1 (2.17%)	2 (4.35%)
CC15	t084	2 (4.35%)	0 (0%)
	t018	1 (2.17%)	0 (0%)
	t306	1 (2.17%)	0 (0%)
	t8786	1 (2.17%)	0 (0%)
CC22	t005	1 (2.17%)	0 (0%)
	t022	0 (0%)	1 (2.17%)
	t310	1 (2.17%)	0 (0%)
	t625	1 (2.17%)	0 (0%)
CC30	t012	6 (13.04%)	0 (0%)
	t019	1 (2.17%)	0 (0%)
	t273	1 (2.17%)	0 (0%)
	t584	1 (2.17%)	0 (0%)
	t8831	1 (2.17%)	0 (0%)
CC45	t015	5 (10.90%)	1 (2.17%)
	t026	1 (2.17%)	0 (0%)
	t040	1 (2.17%)	0 (0%)
	t050	1 (2.17%)	0 (0%)
	t073	1 (2.17%)	0 (0%)
	t339	1 (2.17%)	0 (0%)
	t620	1 (2.17%)	0 (0%)
	t1689	1 (2.17%)	0 (0%)
	t2239	1 (2.17%)	0 (0%)
CC78	t8863	1 (2.17%)	0 (0%)
CC97	t267	3 (6.62%)	0 (0%)
	t131	1 (2.17%)	0 (0%)
	t8831	1 (2.17%)	0 (0%)
CC101	t4044	1 (2.17%)	0 (0%)
CC398	t011	0 (0%)	1 (2.17%)
	t571	1 (2.17%)	0 (0%)
unknown	t078	1 (2.17%)	0 (0%)

Clonal Complex Affiliation

Upon application of the original MA evaluation software (Iconoclust, Alere Technologies), isolates could be assigned to MLST clonal complexes (CCs) based on the hybridization profiles, except for two untypable MSSA isolates (S19, S27) (Figure 5.1B).

The MRSA isolates clustered into only five different CCs, while MA analysis of MSSA revealed twelve different CCs. MRSA isolates were dominated by CC5 $(41, 89.1\%)$ whereas the predominant MSSA types were found to be CC45 $(12, 28.6\%)$ and CC30$(10, 23.8\%)$. Isolates of CC5, CC8, CC22, CC45 and CC398 were found both in the MRSA and the MSSA group, whereas CC30, CC15, CC97, CC7, CC1, CC78 and CC101 were present only in the MSSA group. CCs attributed to the MRSA group only were not found.

Analysis of Gene Equipment

Microarray results of MRSA and MSSA isolates were analysed for individual genes associated with e.g. antibiotic susceptibility, toxin production, adhesion and immune evasion. An overview of the most relevant genes in the investigated isolate cohort was provided for MRSA as compared to MSSA (Figure A.1) Appendix A.

Genes respectively gene components which were not detected in any cohort isolate were not displayed *(ermB, mefA, mph(C), vat(A), vat(B), vga, aphA3, sat, dfrS1, far1, cat, fexA, cfr, vanA/B/C, mercury resistance locus, qacA/C, seb,sef, she, seq, PVL, lukM, etB, edinA/D, splE, vwb, Q2YUB3) as well as allelic variants (vga, lukF, lukS, lukY, hlIII, aur, map, sdrC, sdrD, vwb, sasG, isaB, mprF, ImrP).* For more detailed analysis of selected gene profiles of individual isolates we refer to the supporting information (Table S1). Appendix A

***Agr*-typing**

All CC5 isolates $(n = 41, 89.13\%)$ affiliated with *agr*II (accessory gene regulator type II). The remaining 5 MRSA isolates of CC8, CC22, CC45, CC398 (10.9%) as well as MSSA of CC7, CC22, CC45, CC97, CC101, CC398 $(n = 26, 52.2\%)$ were associated with *agr*I, 12 MSSA isolates of CC1, CC30, CC78 with *agr*III (26.1%) and 7 isolates of CC5 and CC15 with agrII (15.2%). The *agr* type of three MSSA isolates could not be determined using MA.

SCC*mec* Typing

SCC*mec* types were identified based on hybridization patterns. Corresponding to the predominant clonal complex of the MRSA isolates all except four isolates of CC5 (37 of 41, 90.2%) comprised a SCC*mec*-cassette of type II. Isolates of the CC8 (n = 2), CC22 (n = 1), CC45 (n = 1) and one isolate of CC5 harbored the SCC*mec* type IV while the CC398 isolate were characterized by SCC*mec* type V. The SCCmec types of three isolates could not be determined by MA. Resistance Genes MRSA isolates were defined and characterized by the detection of mecA in the SCCmec cassette. 39 MRSA isolates (84.8%) and also 29 (63.0%)MSSA isolates were positive for the $\beta - lactamase$ operon $(blaZ, blaI, blaR)$. 43 (93.5%) MRSA yet only 20 (43.5%) MSSA isolates carried fosB, a putative marker for fosfomycin and bleomycin resistance $(p < 0.001)$; the detection of the *fos*B gene was limited to CC5, CC8, CC15, CC30, CC101. The macrolide, lincosamide and streptrogramin (MLS B) resistance gene *erm*A was detected with significantly higher rates in the MRSA (41, 89.1%) as compared to the MSSA group (3, 6.5%) $(p < 0.001)$. Only one (2.2%) MSSA isolate was positive for *erm*C. The aminoglycoside resistance gene aadD was detected more frequently in MRSA (27, 58.7%) than in MSSA isolates (1, 2.2%) $(p < 0.001)$. Most isolates (84/92 [91.3%]) carried the unspecific efflux pump gene (*sdr*M, formally *tetEfflux*) which was equally distributed among MSSA and MRSA isolates. The tetracycline resistance gene tet(K) was detected in only one MRSA (2.2%) and two MSSA (4.3%) isolates, respectively.

Virulence Genes

Panton-Valentine leukocidin *(pvl)* genes (*luk*F/S-PV) were not detected in the total study cohort. Only 9/46 (19.6%) MSSA isolates were tst1 (toxic shock syndrome toxin) positive, most of them clustering into CC30 (8,17.4%). The genetically linked leukocidin components (*luk*D and *luk*E as well as *luk*S, *luk*F and *hlg*A) were found more frequently in MRSA than in MSSA ($p < 0.001$).

Among the haemolysin gene family, high abundance was detected among MRSA and MSSA for *hla, hlb, hld* and *hl*III, whereas differences between groups were detected for *hlb* ($p < 0.001$).

The immune evasion gene cluster of *sak* (staphylokinase), *chp* (chemotaxis-inhibiting protein), or *scn* (staphylococcal complement inhibitor) was abundantly found both in the MRSA and the MSSA group. Hybridization signals for exfoliative toxin *et*A, *et*B, *et*D and epidermal cell differentiation inhibitor *edin*A, *edin*B, *edin*C genes were detected only in a minority of strains.

The enterotoxin gene cluster (*egc* comprising *seg, sei, sem, sen, seo, seu*) was frequently identified both in MRSA (43/46, 93.5%) and MSSA (29/46, 63%) ($p < 0.001$), yet, the gene cluster was restricted to isolates of CC5, CC22, CC30, CC45. Enterotoxin genes *sea, sed, sej* and *ser* were significantly more frequent in the MRSA group while all isolates were negative for *seb, sef, sek* and *seq* . Interestingly, the 16 isolates of CC7, CC15, CC78, CC97, CC101 and CC398 (one MRSA and 15 MSSA) did not contain any hybridization signal for enterotoxin genes.

The serineprotease genes, *spl*A and *slp*B, were predominantly found in the MRSA group ($p < 0.001$), and this gene cluster was restricted to clonal complexes CC1, CC5, CC7, CC8, CC15 and CC97. The aureolysin gene (aur) was detected in 43 MRSA (93.5%) and 30 MSSA isolates (65.2%) ($p < 0.001$). Other protease genes such as sspA (glutamylendopeptidase), *ssp*B and *ssp*P (staphopain B and A) were detected in the entirety of isolates tested. The ACME gene cluster, which had been brought to attention during analysis of *ca*MRSA outbreak strains, was found in our population in the ST5-MRSA-II group (3, 6.5%).

Microbial surface components recognizing adhesive matrix molecule genes (MSCRAMM) comprising *cna* (collagen-binding adhesin), *sas*G (*S. aureus* surface protein G), *vwb* (van Willebrand factor binding protein) and *fib* (fibrinogen binding protein) are abundantly expressed, however, with higher proportions of *cna* positive isolates in the MSSA group, and higher rates of *fib*, *sas*G and *vwb* in the MRSA group. Other MSCRAMM genes such *bbp* (bone sialoprotein-binding protein), *clf*A (clumping factor gene A), *clf*B (clumping factor gene B), *ebh* (cell wall associated fibronectin-binding protein), *eno* (enolase binding protein), *ebp*S (cell surface elastin binding protein), *fnb*A (fibronectin-binding protein A) and *sdr*C (serine aspartate repeat fibrinogen binding protein) were found in the majority of strains without clear association to the methicillin resistance profiles.

As expected, the most obvious genetic differences in the highly abundant CC5 MRSA group (*bla*-operon, *aad*D, *sea, sed, sej, ser, hlb* and *chp*) were associated with altered mobile genetic elements. More detailed characteristics of individual isolate in respect to *spa*-type, repeat succession, CC, SCC*mec*-type, *agr*-type, toxin profile, resistance profile, strain assignment and relation analysed by hierarchical cluster dendrogram was shown in the supporting information (Figure A.1) Appendix A.

Microarray and spa-type Based Subclassification of CC5 Isolates

Most MRSA isolates were attributed to a genetic group of healthcare associated strains

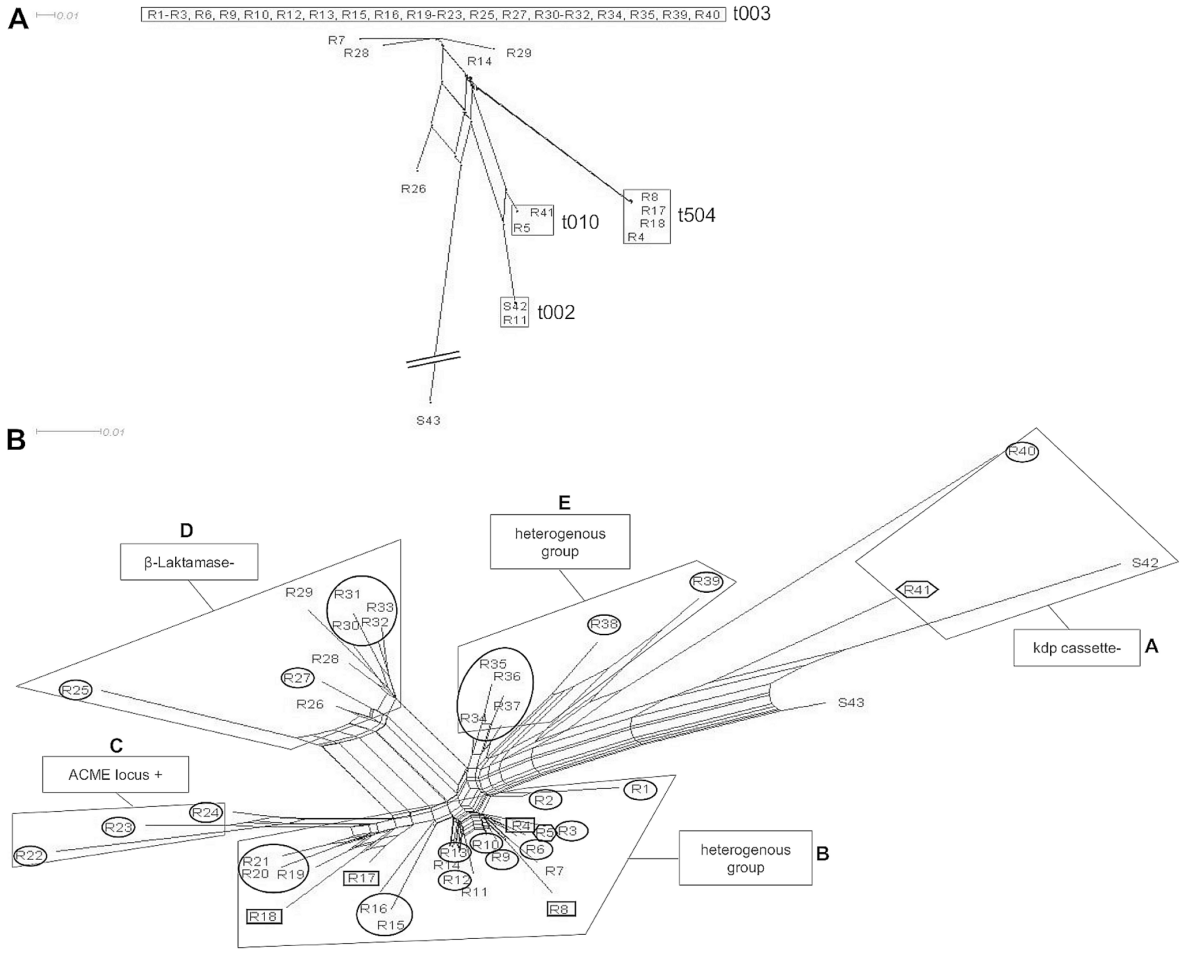


FIGURE 5.2: Subclassification analysis of 41 MRSA (R1R41) and two MSSA (S42, S43) of CC5. (A) Splits graph based on cost distance matrix computed by Ridom StaphType software. (B) Splits graph based on MA hybridization profiles. Characteristic gene profiles for isolate cluster assignment were arbitrarily stated into group A-E. The most common MRSA spa-types t003 (circles), t504 (quadrates) and t010 (hexagons) were highlighed. doi:10.1371/journal.pone.0052487.g002

clustering into the CC5 (41, 89.1%). Except for two isolates of unidentified strain assignment, all isolates of CC5 referred to ST5-MRSA-II. This phylogenetically related and epidemiologically important CC5 was then selected for more detailed subtyping using MA hybridization as compared to classical *spa*-typing. A more detailed subtyping of spa-sequence data beyond the *spa*- type level was not possible as was demonstrated by splits graph distance matrix analysis (Figure 5.2A).

Using the standard IdentiBAC MA software, subtyping of the MA results was not straight-forward. Instead, three alternative bioinformatics methods were found to be very helpful in subdividing genetically related strains by analysis of comprehensive genetic signatures determined by the MA. Results obtained by splits graph analysis (Figure 5.2B), cluster analysis using dendrograms (Figure 5.3), and principal component analysis (PCA) based on MA hybridization signals were evaluated (Figure 5.4). Splits graph of the MA results allowed subclassification of the 41 CC5 isolates into 5 different clusters (A- E), including subclassification of *spa*-type t003 and of both t010 isolates. Interestingly the t504 isolates with regional cumulation clustered exclusively into the subgroup

spa-type	MA PCA cluster (arbitrary group)	MA Splits graph cluster (arbitrary group)	MA Cluster dendrogram (arbitrary group)	Microarray Cluster dendrogram
t003	I d	C	I	R22
t003	I d	C	I	R24
t003	I d	C	I	R23
t003	I a	E	II	R39
t003	I a	E	II	R38
t003	I b	B	II	R1
t003	I b	E	II	R35
t003	I b	E	II	R36
t003	I b	B	II	R16
t003	I b	E	II	R37
t003	I b	B	II	R2
t003	I b	E	II	R34
t504	I c	B	III	R18
t003	I c	B	III	R19
t003	I c	B	III	R20
t003	I c	B	III	R21
t002	I c	B	III	R11
t504	I c	B	III	R17
t045	I c	B	III	R14
t003	I c	B	III	R13
t003	I c	B	III	R15
t003	I c	B	III	R3
t010	I c	B	III	R5
t003	I c	B	III	R6
t504	I c	B	III	R4
t003	I c	B	III	R12
t003	I c	B	III	R10
t003	I c	B	III	R9
t504	I c	B	III	R8
t887	I c	B	III	R7
t003	II	D	IV	R25
t1079	II	D	IV	R26
t003	II	D	IV	R27
t481	II	D	IV	R29
t003	II	D	IV	R30
t3195	II	D	IV	R28
t003	II	D	IV	R31
t003	II	D	IV	R33
t003	II	D	IV	R32
t003	others	E	V	R40
t010	others	A	V	R41
t002	others	A	V	S42
t493	others	A	V	S43

FIGURE 5.3: CC5 isolates (n = 43) characterized by spa -typing and comprehensive MA subgroup analysis using three different bioinformatic modes (principal component analyses, splits graph and cluster dendrogram). doi:10.1371/journal.pone.0052487.g003

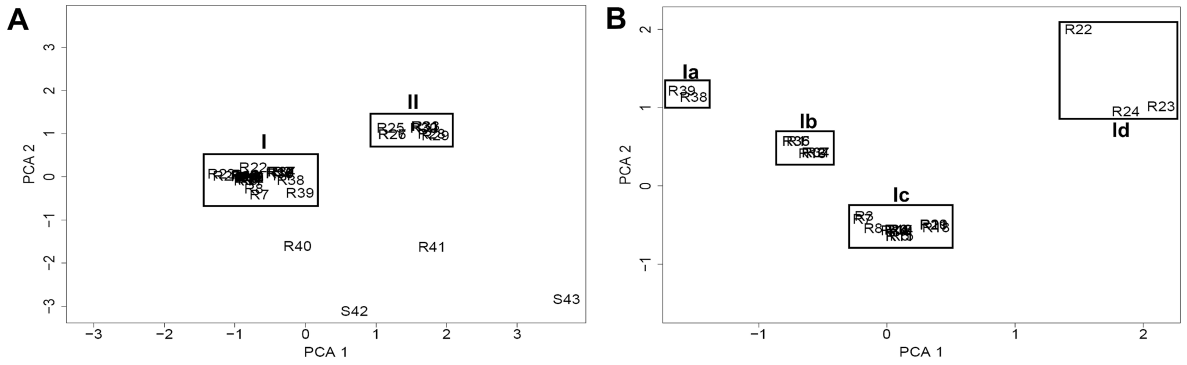


FIGURE 5.4: Principal component cluster analysis (PCA) of 41 MRSA (R1R41) and two MSSA (S42, S43) of CC5. (A) Clustering of the 43 CC5 isolates by PCA as well as (B) subclustering of 30 MRSA CC5 cluster I isolates using a higher resolution PCA plot for in-depth identification of additional subgroups (IaId).
doi:10.1371/journal.pone.0052487.g004

B. Clusters A (*kdp* negative), C (ACME locus positive) and D (b-lactamase negative) were characterized by indicated specific genetic groups, whereas the genetic repertoire of cluster B and E was more heterogeneous. Cluster dendrogram of CC5 isolates revealed similar subclustering as compared to splits graph analysis except for few isolates ($R1, R2, R11, R15, R16, R17$). All CC5 cohort isolates were *agr*II and the majority of CC5 isolates with MRSA resistance profile were SCCmec type II positive strains of the Rhine-Hesse clone (95%). Using PCA, 39 CC5 strains (90.9%) could be discriminated in two major clusters; additionally, four singleton isolates without clustering were found (9.1%) (Figure 5.4 A).

For more detailed information, the predominant cluster I (30 isolates) could be subdivided by focused PCA into four different subclusters (Ia-Id) (Figure 5.4 B) resembling similar subtypes as compared to splits graph and cluster analysis (Figure 5.3).

5.3.1 Discussion

In the present single centre study, the novel IdentiBAC MA platform was applied to the genotypic characterization of matched nasal methicillin sensitive and resistant *S. aureus* isolates collected upon patient admission to a tertiary care university hospital. We could demonstrate that within the colonizing MSSA population tested, a large diversity of CCs was found in contrast to MRSA isolates with limited numbers of CCs and over-representation of CC5/t003. Low lineage diversity in the MRSA in contrast to the MSSA group was found very similarly also in clinical setting e.g. in cystic fibrosis patients [200]. Despite limited number of isolates the IdentiBAC MA revealed significant differences in the genetic repertoire of MRSA vs. MSSA isolates. Genetic differences were found to be distributed among various types of gene families including antimicrobial resistance genes, *agr* types and capsule type. In the present study the MRSA population was characterized by a significantly higher abundance of virulence genes attributed to the leukocidin, enterotoxin, haemolysin, protease and adhesion gene families, whereas only few single virulence genes (*tst*, *ent*L and *cna*) were found more frequently in the MSSA group. Certainly, the genetic profile of the MRSA group was dominated by the genetic repertoire of one single epidemic MRSA clone (Rhine-Hesse); however, it may be also hypothesized that the Rhine-Hesse virulence gene repertoire was relevant for epidemic

spreading of this successful epidemic MRSA clone. Of note, all isolates tested in this study were of commensal nature precluding association of virulence gene equipment with disease, yet, MA may become a regular diagnostic tool if specific clinical features could be associated with virulence gene patterns in subsequent studies.

In this study, it was demonstrated for the first time that evaluation of the raw IdentiBAC MA hybridization data by three independent bioinformatic methods allowed for in-depth phylo- genetic MRSA isolate typing even beyond the prevalent CC5/t003 MRSA genotype. Poor diversity of MRSA with predominance of CC5 isolates could be assumed as a limitation of this study; however, discrimination of these closely related strains is the most important challenge for analysis of healthcare-associated MRSA isolate cohorts obtained from geographically confined studies. In fact, it is the challenge for MA as a new alternative to established typing systems to overcome these limitations. *Spa*-types and MA results were clustered into the same CCs; however, subclustering of the *spa*-types into STs [189] and also MA associated subtyps was not compelling. While genetic signatures of MA allow direct assignment to CCs and STs an assignment to *spa*- types cannot be achieved due to the heterogeneous genetic repertoire in the same *spa*-type. Single run IdentiBAC MA analysis in conjunction with appropriate software tools may now answer detailed questions both of epidemiologic as well as of infection control character.

Splits graph analysis by neighbor joining clustering, cluster dendrogram using hierarchical agglomerative clustering and also principal component analysis (PCA) formed very similar sub- groups of the closely related CC5 isolates. In general, for more detailed strain assignment it has to be amended that a clearcut nomenclature discriminating strains and clones is still missing. In the present study, the CC5 subgroups characterized by a different lineage specific accessory gene repertoire were arbitrarily named group A-E. These predominant subgroups differed for specific gene families encoding b-lactamase resistance *(blaZ/blaI/blaR)* [77], the arginine catabolic mobile element (ACME) [46], the K^+ -transporting ATPase A-C chain, or the sensor histidine kinase, i.e. the *kdp* operon [209]. ACME positive ST5-MRSA-II isolates have been identified before also in Hong Kong and USA (California) which could be the base for new clone/substrain assignment by MA analysis. MRSA strains of the same CC can be attributed to characterized epidemic strains based on the presence/absence of characteristic genes. Thereby, the highly abundant toxic shock gene (*tst*) negative ST5-MRSA-II isolates were identified as Rhine-Hesse clone [136] whereas the CC8- MRSA-IV isolates were attributed to the Lyon clone [112] due to their carriage of enterotoxin A (sea) with or without *sed/sej/ser*. The *tst* positive New-York Japan clone [96] of ST5-MRSA-II.

was not detected in our population. By implementation of MA into routine diagnostics more detailed subtyping with elaborate techniques as e.g. whole genome sequencing can be restricted to few closely related isolates with identical MA profiles clustering in the same genetic subgroup. Differences in characteristic gene families could result in altered metabolism and biologic activity.

However, there is still limited evidence that genetically different subgroups may act differently according to *S. aureus* virulence in vivo. Additionally, also single nucleotide mutations beyond the resolution of the MA may influence the biologic behaviour of *S. aureus* strains which remains undetectable by MA [213]. Correlation between genotypic variants and clinical pheno- type remains to be confirmed in future clinical studies.

While splits graph and cluster dendrogram evaluation are abundantly used for phylogenetic analysis [40], PCA is a dimension reduction model becoming popular in recent

years for genome-wide association studies. Thereby most of the original variability in the data can be retained without organizing them in a hierarchical format.

Comparing the three independent bioinformatic methods, a very similar sub-clustering of closely related CC5 isolates was demonstrated although each model may have its specific strengths for clinical application [149]. The optimal choice between the three methods may indeed depend on the number of samples to be visualized and on the degree of diversity. For example, PCA enables a direct simple overview of an almost unlimited amount of isolates as shown here in the 2-dimensional graph. However, simple assignment of each point in the graph to the corresponding isolate is difficult in the case of densely overlapping samples. On the other hand, cluster dendrogram analysis reveals a more detailed isolate relationship with direct assignment of each isolate to the corresponding subgroup. Yet, this representation is most useful for sample sizes of less than a hundred. In the present case, splits graph analysis appeared to be most appropriate for diversity analysis during routine diagnostics due to ease-of-applicability, open- source software tools and direct assignment of each isolate to the branched subgroups in the 2-dimensional graph. For future application of MA as an internationally accepted diagnostic tool it is important that a common standardized database-associated software tool is implemented independent of universally applicable bioinformatic tools investigated in the present study.

In conclusion, the present matched control study demonstrated a high genetic diversity for MSSA, either directly by spa-typing or by MA. However, differentiation of the predominant epidemic CC5 MRSA isolates was limited for *spa*-typing whereas detailed subtyping was achieved by bioinformatic-assisted MA analysis. The IdentiBAC MA could fulfil a number of criteria required for a new standard test for *S. aureus* typing including standardisation, ease of performance, low turn-around time ($< 24hours$), appropriate costs and superiority to established typing methods as was shown here for *spa*-typing. Based on the IdentiBAC MA concept, and as goal for the future development, standardized and easily applicable software tools based on the bioinformatic approaches with set highly differentiated strain assignment would then allow for comprehensive strain differentiation and global data exchange.

Chapter 6

Community-Acquired *Staphylococcus aureus* Isolates From Various Sub-Saharan African and German regions: Clonal Cluster Analysis Reveals Significant Differences by Geographical Origin and Clinical Significance

This chapter is based on a manuscript resulting from a collaborative project funded through the DFG-Africa initiative. Our project partners Prof. Dr. med. Mathias Herrmann, Prof. Dr. med. Lutz von Müller and Ulla Ruffing were responsible for sample collection and generation of experimental data. My task was to process the data from the Alere IdentiBAC chip. The manuscript was written in collaboration with Prof. Dr. med. Mathias Herrmann and Ulla Ruffing.

In developing and emerging countries limitation of molecular, pathogenicity and epidemiologic data of pathogens as *S. aureus* is a great challenge [33, 73, 78]. Many African studies demonstrate the frequency, resistance profile and higher mortality associated with *S. aureus* diseases in comparison to developed countries [148]. But there are only a limited number of publications available of *S. aureus* prevalence, genotype as well as of PVL prevalence [2, 69]. This is the situation particularly for the Central African region in comparison to other African regions where more data becomes available.

Recent studies demonstrate that *S. aureus* is one of the most often isolated bacteria of infections in Sub-Saharan Africa [6, 144], that community-acquired *S. aureus* bacteremia (9.5%) could be seen more often than meningococcal sepsis (1%) and that it could be identified as one of the leading causes of bacteria sepsis in Nigeria [151, 196, 166]. Altogether publications with the focus on community acquired CA-MRSA and CA-MSSA in Africa including one multicentre study comprises only small isolate numbers

[57, 69, 175, 163]. Available African and international multicenter studies including a limited number of African isolates are performed using different molecular techniques as spa-typing, Multilocus sequence typing (MLST) and pulse-field-gel-electrophoresis (PFGE). A multicenter study comparing molecular and epidemiologic data based on the same identical technique of different African and European isolates of equal numbers for comparison of local and international clone distribution was not performed till yet.

Community acquired MRSA is defined as any infection diagnosed in an outpatient or within 48h of hospital admission not fulfilling any of the following risk factors of HA-MRSA: haemodialysis, surgery, and residence in a long term care facility or treatment during the last year, presence of a permanent catheter or percutaneous device or previous isolation of MRSA. The hospital environment still is a risk factor for *S. aureus* carriage and nosocomial infections but an increasing number of *S. aureus* infections are caused by community-associated strains [34, 157]. This means that the known differences of the established clonal structures are burring too. Differences of clonal lineages of *S. aureus* could be influenced by their localization isolate of an asymptomatic healthy volunteer or from a patient infection site as well as by their geographic origin [74]. But little is known about the population structure and geographical abundance of methicillin susceptible *S. aureus* (MSSA) as genetic MRSA reservoir in the African as well as in the German community.

In this multicenter study three African and three German study sites collected 1200 African and German community acquired *S. aureus* with the aim of the analysis and comparison of their molecular characteristics and clonal structure based on one identical molecular method (DNA microarray, Alere). Thereby carriage isolates of healthy volunteers and infection site isolates of patients without previous clinical contact for further analysis of isolate source dependent gene associations were discriminated.

6.1 Material and Methods

Study design and participants. In this prospective cohort study of the German-African network on staphylococci and staphylococcal disease (DFG PAK 296) 1200 community-associated isolates were collected in three African (Lambarn, Gabon; Dares-Salaam,Tanzania; Ifakara, Mozambique) and three German study sites (Homburg, Saarland; Freiburg, Baden-Wrttemberg; Mnster, Nordrhein-Westfalen) in accordance to the predefined case-related-forms (CRFs) of the StaphNet consortium to exclude hospital acquired *S. aureus* strains. Every study site collected 100 isolates of healthy volunteers and 100 isolates of clinical infection sites of people without previous clinical contact during the last half year. The clinical data were compiled with the CRFs and collected in one database for all 1200 cases.

All isolates were assigned an unique strain identification encoding the study site LG= Lambarn, Gabon, MM= Manhica, Mozambique, IT= Ifakara, Tanzania, HS = Homburg, Saarland, MW= Münster, Westfalen, FR= Freiburg, Baden-Württemberg, the strain origin N= nasal, B= blood culture, O= wound infection and others and its specific number as e. g. IT-N075.

Ethics approval was obtained from the Ministry of Health and social Welfare of Tanzania, Institutional Ethics Committee of the Medical Research Unit of the International Foundation of Dr. Albert-Schweitzer Hospital (Lambarn, Gabon), Comit nationale de Biotica para a saud (Manhia, Mozambique), ethics committee of the medical association

and medical department of the Westfälische Wilhelms-University (Münster, Nordrhein-Westfalen), ethics committee of the medical department of the Albert-Ludwigs-University (Freiburg, Baden-Wrttemberg) and the ethics committee of the medical association of the Saarland (Homburg, Saarland). A written informant consent was received from all study subjects or their legal guardians.

Isolate collection and microbiological methods. Nasal swabs were collected using nasal swabs of healthy volunteers with no clinical contact during the last half year. Strains from infections sites were collected in the different health institutes of the six study sites according to standard procedure and methods. Nasal swabs and infection sites strains were cultured using standard methods. Isolates were identified by colony morphology on sheep blood agar, catalase test, latex test $Pastorex^{[TM]}$ Staph-Plus-Latex Test, Bio-Rad, Marnes-la-Coquette/Frankreich) and identity confirmation by Maldi-TOF (BRUKER Daltonics).

According to the Kirby-Bauer-method disk diffusion tests were performed for penicillin, cefoxitin, tetracycline, erythromycin, clindamycin, gentamycin, chloramphenicol and cotrimoxazol, MICs were determined by E-Test for cefoxitin, clindamycin, linezolid, vancomycin, daptomycin and tigecyclin (CLSI, M100-S16, 2006); furthermore inducible clindamycin resistance was carried out by performing D-test using (CLSI, M100-S16, 2006) criteria.

DNA microarray-based genotyping and MLST. DNA extraction and hybridization to the IdentiBAC MA (Alere Technologies GmbH, Jena, Germany) was performed as described in the manufacturers instructions in combination with DNeasy blood and Tissue kit (Qiagen, Hilden, Germany) [139, 137]. The test principle is based on a linear multiplex primer elongation using one primer for every single target and DNA labeling by incorporation of biotin-16-dUTP in the approximately 40-fold DNA amplification. DNA hybridization microarray probes were washed and then horseradish-peroxidase-streptavidin precipitation reaction was performed resulting in visible grey spots in case of a positive reaction. Spot signals were recorded and automatically analyzed using the designated ArrayMate reader and the corresponding software (Iconoclust, Alere Technologies) [139]. In conjunction with the Iconoclust analysis, array profiles are attributed to a specific clonal complex (CC) and sequence type (ST) based on a proprietary algorithm provided by the manufacturer. Similarly, SCCmec types are attributed as a result of array signals obtained. Multilocus sequence typing (MLST) was performed for samples without CC assignment by the DNA microarray as published previously [58].

CC assignment confirmation and Statistics. Correctness of the CC identification by DNA-MA was confirmed by next generation sequencing (NGS) of 160 representative samples. DNA purification (MagAttract HMW DNA Kit (Qiagen, Hilden, Germany) and NGS (MiSeq, Illumina, San Diego, USA) was performed according to manufacturers instructions. Obtained reads were de novo assembled using the velvet assembler implemented in the software SeqSphere+ (version 2.0, Ridom GmbH, Mnster, Germany) with a minimum coverage of 5 and an aspirated mean coverage of 100. Short reads <200 nucleotides were excluded. The multilocus sequences typing sequence types (MLST ST) were inferred from the data according to a WGS adapted scheme of Enrights method [58] [PMID:10698988] including an up-to-date comparison with the online MLST database www.mlst.net using SeqSphere+.

Exemplars were defined by Affinity propagation. Affinity propagation is a clustering algorithm that has as input measures of similarity between pairs of data points (e. g. isolates with DNA microarray data) and simultaneously considers the whole data points

as potential exemplars out of the collection. An exemplar is a member (e. g. isolate) of the data input (isolate and its genotype data) which is representative for a cluster (e.g. group of isolates).

Principal component analysis (PCA) was performed to represent the isolates genotype in a two-dimensional graphical space. PCA reduce the dimensionality of the MA data and identified groups of correlated variables. Multivariate analysis (Kolmogorov-Smirnoff test) was used to determine genotypic differences of isolate clusters defined by (PCA) (Figure 6.4).

All comparisons were statistically analyzed by Chi-Square test using Graph pad (online tool); differences yielding statistical significance ($p < 0.05$) were annotated (not shown). A correction for multiple testing was not employed. Multivariate and principal component analysis was performed with the software R, version 3.2.0.

Silhouette analysis. The silhouette plot based on KMeans clustering displays a measure of how close each point (sample) in one cluster is to points (isolates) in the neighbouring clusters and thus provides a way to assess parameters like a number of clusters visually. It is a method to measure the strength of clusters or how well one element was clustered. The measure has a range of [-1, 1]. The silhouette analysis was performed to determine the number of different isolate clusters in the PCA. Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighbouring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighbouring clusters. Negative values indicate that those samples might have been assigned to the wrong cluster.

6.2 Results

Patients and isolates characteristics. A total of 600 nasal and 600 clinical associated isolates were included; each study site collected 100 nasal isolates from healthy volunteers and 100 isolates from patients from clinically significant specimens, positive samples taken $\leq 48h$ of admission (including at transferring hospital).

The patients characteristics are summarized in Table 1.

Table 6.1: Differences of spa-types and clonal complexes in MSSA and MRSA isolates.

	Africa	**Germany Commensal**	**P**	**Africa**	**Germany clinical**	**P**
Total	300	300		300	300	
From blood culture	na	na		29	50	
From clinically material other than blood	na	na		271	250	
Median age (range)	17 (0-71)	23 (0-89)		0-71	0-98	
Female	167 (56%)	150 (50%)	0.1907	144 (48%)	112 (37%)	0.0104
Patient history						
- hospitalization last 6 months	0	14 (5%)	0.0001	29 (10%)	128 (43%)	0.0001
- close health care contact last 30 days	na	na	na	36 (12%)	94 (31%)	0.0001
- nursing home	0	0	1	2 (>1%)	2	1
- tuberculosis last 6 months:	na	na	na	4 (1%)	0	0.1237
- antituberculous drugs last 4 weeks	0	0	1	1 (>1%)	0	1
- including rifampin	0	0	1	1 (>1%)	0	1
- antibiotics last 4 weeks	0	0	1	55 (18%)	78 (26%)	0.0304
McCabe-Jackson underlying disease prognosis						
- rapidly fatal	0	0		10	11	
- fatal within the next few years	0	1		5	47	
- not expected to be fatal within the next 4 years	5	8		3	110	
Comorbidities						
- HIV infection	14	0	0.0001	25	0	0.0001
- AIDS	0	0	1	8	0	0.0075
- Myocardial infarction/coronary heart disease	2	4	0.6859	0	37	0.0001
- Congestive heart failure	0	3	0.2487	0	26	0.0001
- Peripheral vascular disease	0	4	0.1237	0	65	0.0001
- Cerebrovascular disease	0	0	1	0	25	0.0001
- Dementia	0	0	1	0	7	0.0151
- Chronic obstructive pulmonary disease	2	0	0.4992	2	17	0.0006

Continued

	Africa	Germany Commensal	P	Africa	Germany clinical	P
- Connective tissue disease	0	4	0.1237	0	31	0.0001
- Peptic ulcer disease	8	1		2	11	0.0211
- Mild liver disease	0	3	0.2487	0	22	0.0001
OR moderate-severe liver disease	0	0	1	0	7	0.0151
- Diabetes mellitus	0	8	0.0075	4	54	0.0001
OR diabetes mellitus with organ damage	0	1	1	0	15	0.0001
- Hemiplegia	1	0	1	2	4	0.6859
- Moderate-severe renal disease	0	1	1	0	25	0.0001
- Any tumour (within last 5 years)	0	0	1	0	50	0.0001
Lymphoma	0	0	1	0	5	0.0615
Leukemia	0	0	1	0	4	0.1237
Metastatic solid tumor	0	0	1	0	20 0.0001	
SAB risk factors						
- IVDA	0	0	1	0	7	0.0151
- vascular catheter	0	0	1	1	21	0.0001
- vascular foreign body	0	6	0.0305	0	27	0.0001
- other foreign body	17	19	0.8638	2	70	0.0001
Resistance phenotype						
Penicillin	271	181	<0.0001	289	203/279	
Cefoxitin	7	2	0.1764	10	21	0.0636
Tetracycline	106 (35%)	4 (1%)	<0.0001	145 (48%)	17 (6%)	0.0001
Erythromycin	61 (20%)	46 (16%)	0.1352	56 (19%)	59 (20%)	0.8357
Gentamicin	15	1	0.0004	3	8	0.2222
Choramphenicol	7	1	0.0685	13	3	0.0196
Cotrimoxazole	42 (14%)	0 (<1%)	<0.0001	58 (19%)	4 (1%)	0.0001
Inducible clindamycin resistance	48 (16%)	40/207		47 (16%)	14/153	
Inducible clindamycin resistance	48 (16%)	-19%		47 (16%)		
	Africa	Germany		Africa	Germany	

Continued

	Africa	**Germany Commensal**	**P**	**Africa**	**Germany clinical**	**P**
	O (N=271)	O (N=250)	p-value	B (N=29)	B (n=50)	p-value
Severe systemic disease						
Severe sepsis	3 (1%)	3 (1%)	1	15 (52%)	16 (32%)	1
Septic shock	1	1	1	1 (3%)	4 (8%)	0.3729
Clinical site(s) of infection						
- superficial [skin and skin structure]	184 (68%)	133 (53%)	<0.0001	9 (31%)	20 (40%)	0.0553
- deep skin abscess	78 (29%)	37 (15%)	<0.0001	4	3	1
- other (deep) sites:						
- bone	0	20	<0.0001	0	2	0.4992
- joint	2	12	0.0120	1	3	0.6237
- thigh	1	0	1	1	1	1
- muscle: regions other than thigh	0	5	0.0615	1	0	1
- fasciitis	0	1	1	0	0	1
- respiratory tract/lungs incl. pleura	1	16	0.0002	5	2	0.4504
- heart/heart valve	0	0	1	0	5	0.0615
- CSF/brain	1	0	1	0	0	1
- urinary tract	0	4	0.1237	0	4	0.1237
New metastatic lesions						
- any	na	na	na	2 (7%)	30 (60%)	0.0001
- bone	na	na	na	0	6	0.0305
- joint	na	na	na	0	1	1
- deep skin abscess	na	na	na	1	2	1
- thigh	na	na	na	0	0	1
- other muscle	na	na	na	0	4	0.1237
- fasciities	na	na	na	0	0	1
- respiratory tract	na	na	na	1	1	1
- heart	na	na	na	0	6	0.0305
- brain	na	na	na	0	1	1

Continued

	Africa	Germany Commensal	P	Africa	Germany clinical	P
- urinary tract	na	na	na	1	0	1
- other	na	na	na	0	6	0.0305
Patient admitted because of SAB within 14 days				28 (97%)	43 (86%)	0.0762
Patient admitted because of SAB within 3 days	47 (17%)	138 (55%)	<0.0001			
Patient died within 14 days	0	3 (1%)	0.2487	1 (3%)	6 (12%)	0.1228
Antibiotic therapy day 1	245 (90%)	127 (51%)	<0.0001	29 (100%)	43 (86%)	0.1019
Antibiotic therapy day 2	243 (90%)	127 (51%)	<0.0001	29 (100%)	44 (88%)	0.0798
Antibiotic therapy day 3	242 (89%)	129 (52%)	<0.0001	29 (100%)	45 (90%)	0.0620
Antibiotic therapy day 5	na	na	na	27 (93%)	45 (90%)	0.0321
Antibiotic therapy day 7	na	na	na	26 (90%)	44 (88%)	0.0301
Antibiotic therapy day 10	na	na	na	18 (62%)	41 (82%)	0.0023
Antibiotic therapy day 14	na	na	na	9 (31%)	37 (74%)	0.0001
Type of intervention						
I and D surgery within 3 days	168 (62%)	66 (26%)	<0.0001	3 (10%)	11 (22%)	0.0545
I and D surgery days 4-7	na	na	na	1	6	0.1228
I and D surgery days 8-14	na	na	na	1	6	0.1228
Vascular catheter removal within 3 days	na	na	na	2	8	0.1064
Vascular catheter removal days 4-7	na	na	na	3	1	0.6237
Vascular catheter removal days 8-14	na	na	na	0	3	0.2487
Foreign body removal within 3 days	3	17	0.0022	0	1	1
Foreign body removal days 4-7	na	na	na	1	1	1
Foreign body removal days 8-14	na	na	na	0	1	1
Other surgery within 14 day	na	na	na	1	7	0.0685
Other surgery within 3 days	12	41	<0.0001	na	na	na
Antibiotic therapy day 1	245 (90%)	127 (51%)	<0.0001	29 (100%)	43 (86%)	0.1019
Monotherapy	215 (79%)	90 (36%)	<0.0001	12 (41%)	21 (42%)	0.1510
Antibacterial agents used						
- Pen/Amoxi	75	4	<0.0001	19	0	0.0001

Continued

	Africa	**Germany Commensal**	**P**	**Africa**	**Germany clinical**	**P**
- /	7	9	0.8009	1	10	0.0110
- Oxa	106	9	<0.0001	2	2	1
- Cef1/2	1	43	<0.0001	0	11	0.0009
- Cef3	10	20	0.0903	4	7	0.5450
- Clinda	0	32	<0.0001	0	3	0.2487
- Macrolides	34	2	<0.0001	0	6	0.0305
Antibiotic therapy day 2	243 (90%)	127 (51%)	<0.0001	29 (100%)	44 (88%)	0.0798
Monotherapy	213 (79%)	89 (36%)	<0.0001	11 (38%)	15 (30%)	0.5483
Antibacterial agents used						
- Pen/Amoxi	51	4	<0.0001	18	1	0.2531
- /	7	16	0.0869	1	8	0.0378
- Oxa	101	10	<0.0001	3	7	0.3396
- Cef1/2	1	42	<0.0001	0	10	0.0018
- Cef3	9	21	0.0903	4	8	0.3828
- Clinda	0	30	<0.0001	0	4	0.1237
- Macrolides	34	2	<0.0001	0	4	0.1237
Antibiotic therapy day 3	242 (89%)	129 (52%)	<0.0001	29 (100%)	45 (90%)	0.0620
Monotherapy	212 (78%)	91 (36%)	<0.0001	14 (48%)	16 (32%)	0.8518
Antibacterial agents used						
- Pen/Amoxi	73	4	<0.0001	16	1	0.0002
- /	7	16	0.0869	0	7	0.0151
- Oxa	102	10	<0.0001	3	11	0.545
- Cef1/2	1	49	<0.0001	0	10	0.0018
- Cef3	9	20	0.0903	6	6	1
- Clinda	0	30	<0.0001	0	4	0.1237
- Macrolides	34	2	<0.0001	0	4	0.1237

Of the African and German volunteers with positive nasal cultures, 56% and 50% were female, while in patient cohort providing clinical isolates, 48% and 37% were female in the African and the German cohort, respectively. Clonal complex affiliation. Upon application of the original MA evaluation software (Iconoclust, Alere Technologies) 1193 isolates of the 1200 *Staphylococcus aureus* isolates could be assigned to established 32 clonal complexes (CC) and three sequence types based on the hybridization profiles. Clonal complex assignment with DNA microarray could not be performed for 8 isolates. For these 8 isolates, new sequence types (ST) were identified (ST2370, ST2678, ST2733, ST2735, ST2744) by MLST which were not covered before by known array profiles. In a pilot of 160 selected isolates, NGS was applied in parallel to assess the validity of ST/CCs assignment by the DNA microarray as compared to the current sequence-based gold standard. CC assignment using MA was correct for 154 (96.3%) isolates. Interestingly, strain assignment for CCs could be optimized for three isolates (1.3%) using CC assignment of MA profiles using affinity propagation analysis instead of the original IdentiBAC software.

CC association with geographic *S. aureus* origin or with clinical significance. Except of four CCs (CC80, CC88 in Africa and CC50, CC398 in Germany) all other CCs with a number of at least six isolates were found in Africa as well as in Germany. Isolates of CC5, CC8, CC9, CC25 and CC707 were equally distributed in Africa and Germany, while significant differences for the geographic CC distribution in Africa and Germany were found for 17 of the 40 detected CCs and sequence types (ST). Predominant African CCs were CC1 ($p < 0.0001$), CC6 ($p = 0.002$), CC15 ($p < 0.0001$), CC80 ($p = 0.0002$), CC88 ($p < 0.0001$), CC121 ($p < 0.0001$) and CC152 ($p < 0.0001$). In Germany the most common CCs are CC7 ($p < 0.0001$), CC12 ($p = 0.0002$), CC22 ($p < 0.0001$), CC30 ($p < 0.0001$), CC45 ($p = 0.001$), CC50 ($p = 0.03$), CC59 ($p = 0.02$), CC97 ($p = 0.003$), CC101 ($p = 0.001$) and CC398 ($p < 0.001$) (Figure 1). Additionally CC121 ($p < 0.0001$) and CC152 ($p < 0.0001$) were significantly more often found in isolates of clinical origin, whereas CC45 ($p < 0.0001$), CC101 (p=0.03) and CC707 ($p = 0.03$) where significantly more often identified in nasal isolates (Figure 6.1).

Clinical and nasal origin dependent CC distribution in African and German study sites. While above data describe the CC attribution on overall differences between the clinical/commensal and African/German groups, figure 3 details the proportions of CCs as a function of the clinical significance within the geographic subgroups, as well as the CCs proportion as function of geographic origin within the two clinical significance groups (left panel). In addition, the relative proportions of the different CCs annotated by their respective institution of origin were analyzed (right panel).

When comparing the 600 African strains (left panel, first and third bar), in clinical isolates a significantly larger proportion was found for CC121 (yellow section) ($p < 0.0001$) and CC152 (black)($p < 0.001$), while in nasal isolates a significantly larger proportion of CC8 (orange) ($p < 0.05$) and CC45 (red)($p < 0.0001$) was ascertained. In the group of strains from Germany (left panel, second and fourth bar), a significant larger proportion of CCs could only be found in the group of nasal isolates, i.e. the CC15 complex (blue section) ($p < 0.05$).

When inspecting the differences in CC proportion analysed as a function of the institutional origin (Figure 6.2, right panel, with the African and German institutions in both group of columns depicted from left to right, respectively), the following observations could be made: Within the group of isolates collected in Ifakara (Tanzania), CC121 (yellow) ($p < 0.0001$) and CC152 (black) ($p < 0.0001$) were more predominantly

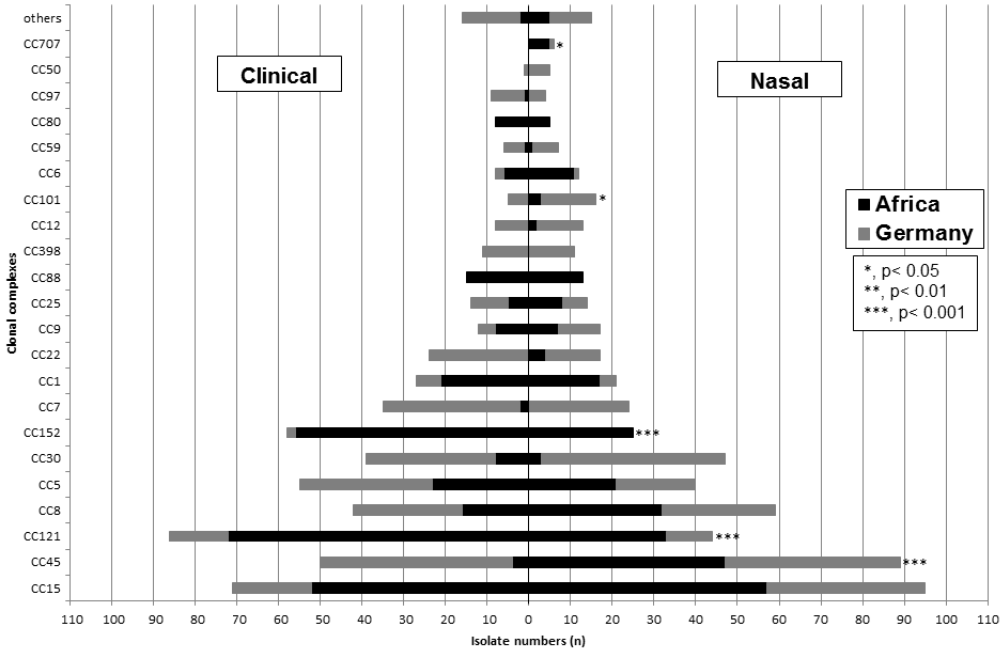


FIGURE 6.1: Prevalence of the 22 most prevalent clonal complexes comparing the nasal or clinical origin. Nasal isolates were collected by nasal swabs. Clinical isolates were of different wound infections or blood cultures. Clonal complexes and sequence types with less than six isolates where taken together as others. The clonal complexes in the y-axis were ordered according to the total number of identified isolates independent of the study site from the bottom to the top. Prevalence of German (grey) and African (black) isolates in the clinical (left side) and nasal group (right side) is shown. Statistical analysis for association of clonal complexes with clinical or nasal isolate origin were performed by Fishers exact test; *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

identified in the group of clinical strains while CC8 (orange) ($p < 0.0001$), while CC15 (blue)($p < 0.5$) and CC45 (red) ($p < 0.0001$) were overrepresented in the group of nasal isolates. In Lambarn, within the group of clinical strains CC1 (purple) ($p < 0.05$) and CC152 isolates (black) ($p < 0.001$) were predominant while CC45 isolates (red) ($p < 0.0001$) were overrepresented by isolates of nasal origin. In Manhiça an overrepresentation of isolates belonging to the CC121 complex (yellow) ($p < 0.01$) could be observed in the group of clinical origin, similarly to Ifakara. Contrasting to the African study sites, the CC distribution of isolates collected in the three German institutions was quite homogenous, both in the overall proportions comparing clinical versus nasal strains as well as with respect of differences within the institutions. Only at the study site in Homburg, a slight but significant association of CC121 (yellow) ($p < 0.0001$) of clinical origin was found, and similarly to Ifakara and Manhiça - CC15 strains (blue) ($p < 0.05$) were more frequently encountered in the nasal isolate group obtained in Freiburg.

Identibac microarray target recognition as a function of geographical origin and clinical significance of *S. aureus* isolates.

***S. aureus* species markers.** The *S. aureus* species markers rrn, gapA, katA, coa,

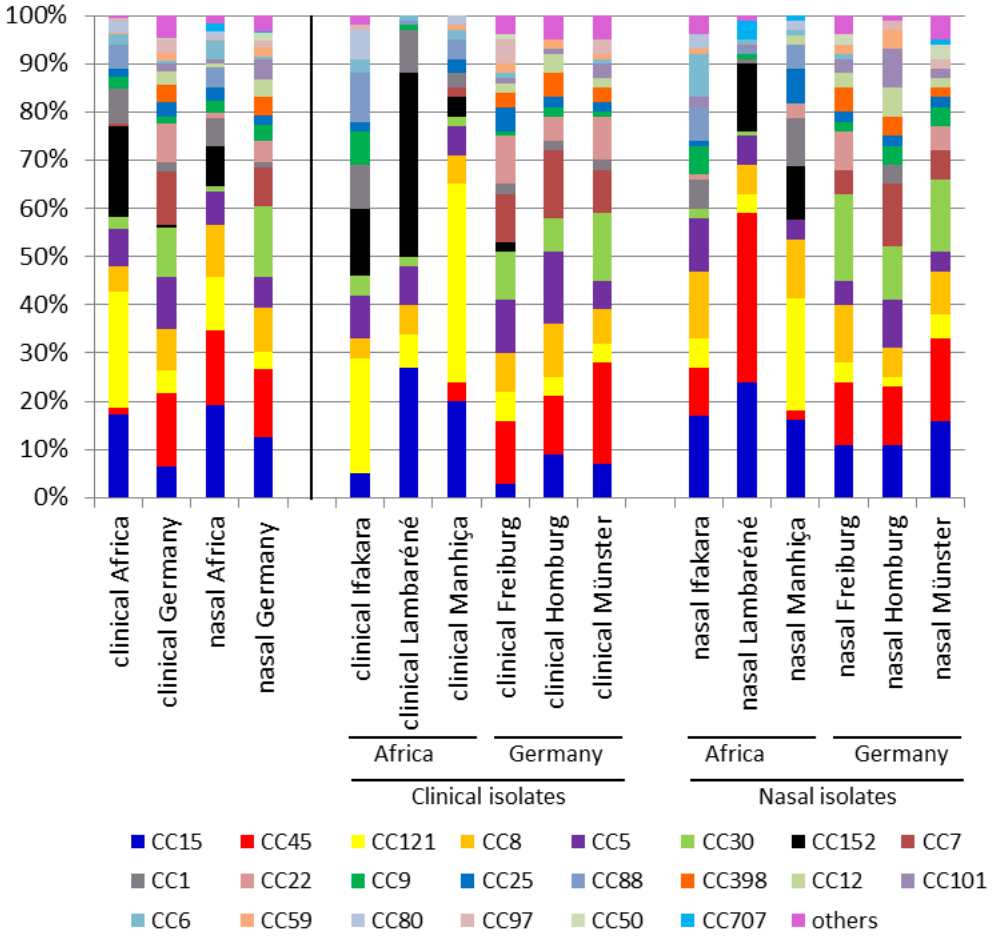


FIGURE 6.2: Relative abundance of the 22 most prevalent clonal complexes in the different study sites and the isolates body localization. Nasal isolates were collected by nasal swabs. Clinical isolates were of different wound infections or blood cultures. Clonal complexes and sequence types with less than six isolates where taken together as others.

nuc1, spa, and sbi were all unanimously (100%) positive in all isolates tested providing for an internal quality control both of the species identification during the isolate recovery as well as for the microarray.

Gene regulators. The accessory gene regulator-I (*agr*I) encoding genes revealed an overrepresentation in German isolates (55% vs 35%, $p < 0.0001$) while the accessory gene regulator-IV (*agr*IV) encoding genes were overrepresented in African isolates (37% vs 6%, $p < 0.0001$). In commensal samples, the capsule type 5 encoding genes were predominant in isolates from African volunteers (43% vs 33%, $p < 0.01$) while capsule type 8 encoding genes were more frequent in isolates obtained from nasal samples (67% vs 56%, $p < 0.01$) in Germany.

Methicillin resistance and SCCmec typing. With respect to the presence or absence of the methicillin resistance gene cassette, mecA, in total 40/1200 (3%) strains were found to carry this gene. MRSA strains were found to be equally distributed in isolates from Africa ($n = 17, 3\%$) and from Germany ($n = 23, 4\%$). Moreover, in Africa, the MRSA isolates were equally distributed between clinical ($n = 9$) and nasal isolates ($n = 8$)

(Table S5, first panel, third gene group), yet in Germany significantly ($p < 0.0001$) more MRSA isolates could be observed in clinical ($n = 21$, 7%) than in nasal isolates ($n = 2$, 1%).

Betalactamase resistance. The beta lactamase resistance operon (blaZ, blaI, blaR) was detected throughout isolates from CCs in Africa and Germany, yet, in African isolates ($p < 0.0001$) it was predominant with more than 90% of the isolates carrying this gene while only approximately 2/3 of the German isolates carried the bla operon. This difference was independent of clinical versus nasal origin (Table S1, fourth gene group), and it was particularly pronounced in CC5 (98% versus 53%, $p < 0.0001$), CC8 (100% versus 74%, $p < 0.0001$)and CC45 (86% versus 55%, $p = 0.0002$). Independent of the isolates geographic origin blaZ, blaI, blaR could be found more often in CC45 nasal isolates (74% versus 52%, $p < 0.01$) while in the CC152 isolates, blaZ could be more frequently (98% vs. 80%, $p < 0.01$) observed in the clinical group.

Other resistance markers. The erythromycin resistance genes ermA were detected in 12 different CCs ($n = 43, 4\%$) and ermC in 21 CCs ($n = 134, 11\%$). In German isolates, ermA was more frequently positive when compared to isolates from Africa (7% vs. 1%, $p < 0.001$), both in clinical (8% vs. 1%, $p < 0.001$) and nasal isolates (5% vs. 0%, $p < 0.001$) while for ermC such higher positivity could be seen in African isolates overall (15% vs. 7%, $p < 0.001$) as well as in the subgroups of clinical (14% vs.8%, $p < 0.05$) or nasal origin (17% versus 6%, $p < 0.001$). The tetracycline resistance gene tetK were found in all CCs in Africa except of CC30. In Germany it could be detected in CC5, CC7, CC8, CC15 and CC121. A significantly larger proportion of tetK and tetM were found in African isolates independent whether they were of clinical or nasal origin (35% vs. 3% for tetK and 8% vs. 1% for tetM, $p < 0.001$). Moreover, in African isolates an association of *tet*K positivity with the CCs CC5 ($p < 0.01$), CC15 ($p < 0.001$), CC30 ($p < 0.001$) and CC45 ($p < 0.001$) was found while *tet*M was significantly associated with CC8 ($p < 0.01$) and CC121 ($p < 0.001$) (Table S1).

Particularly elevated positivity rates were found for the fosfomycin resistance marker fosB, yet, this resistance gene was detected in even a higher proportion in clinical African isolates (62% vs. 57%, $p < 0.01$). Interestingly, the distribution of *fos*B was found to be very heterogeneous. In CC5, CC8, CC15, CC30 and CC121 (almost) 100% of the isolates were positive while all 139 isolates belonging to CC45 were negative. Only isolates of CC1 revealed a difference when comparing African versus German origin (11% versus 60%, $p < 0.01$). Efflux resistance markers (qak) were only rarely found to be positive, and while almost all ($> 95\%$) of isolates were positive for the sdrM gene (encoding a *S. aureus* multidrug efflux pump), the other genes associated with glykopeptide resistance were unanimously found to yield a negative signal.

Toxic shock syndrome toxin. The toxic shock syndrome toxin tst1 was found in 13 different CCs ($n = 103, 12$

Enterotoxins. For sea (staphylococcal enterotoxin A) only CC specific geographic differences were identified. In African isolates, the gene was predominant in CC5 and CC15, while in CC30 it was predominant in German isolates. Overall, *seb* recognition was also predominant in African isolates (19% vs. 8%, $p < 0.001$), in particular in CC5 and CC121 complexes (table S1), yet, independent of the geographic origin this enterotoxin was more often found in clinical isolates(16% vs. 11%, $p < 0.05$). sec and sed were overrepresented in clinical German isolates ($p < 0.001$). A higher predominance of sej ($p < 0.01$), sel ($p < 0.001$) especially in CC1 isolates ($p < 0.001$) and ser ($p < 0.01$) was identified for clinical German in comparison to clinical African samples. In contrast

sek and *seq* were more often in nasal African than in nasal German isolates ($p < 0.05$). The enterotoxin gene cluster egc consisting of the six enterotoxin genes *seg*, *sei*, sem, *sen*, *seo* and *seu* were significantly more often ascertained in clinical German isolates ($p < 0.001$) but CC specific they could be detected more often in African CC8 isolates ($p < 0.001$).

Leukocidins. The leucocidin genes *luk*D, *luk*E, *luk*Y were predominant in African isolates (71% vs. 55%, 73% vs. 54%, 89% vs. 72%, respectively, $p < 0.0001$) with *luk*D and *luk*E particularly pronounced in CC1. *luk*X was most frequently found in German isolates (98% vs. 86%, $p < 0.0001$). Moreover, both *luk*X ($p = 0.0002$) and *luk*Y ($p = 0.0008$) were predominant in clinical isolates. The major difference between African and German isolates, however, was ascertained when testing for the PVL encoding genes: Overall, *luk*F-PV and *luk*S-PV was found in 287 (24

Immune evasion cluster converting phage genes. The genes sak encoding Staphylokinase, chp, encoding the protein *CHIPS and scn, encoding the protein SCIN are genes of the immune evasion cluster.* sak and scn were recognized in the large majority (80% and above) of isolates, while recognition of chp was only seen in approximately one half of the isolates. Moreover, sak was found predominant in clinical isolates (82% vs. 76

Exfoliatin genes. *et*A and *et*B could be detected in nine ($etA; n = 63, 19\%$) and five ($etB; n = 28, 0.02\%$) different CCs. Both toxin genes *et*A ($p < 0.01$) and *et*B ($p < 0.05$) were overrepresented in clinical CC15 and CC121 isolates independent of the isolates geographic origin as well as etA was significantly more often identified in German CC121 isolates ($p < 0.01$). *ACME gene cluster.* The arginine deiminase gene cluster was only rarely identified, in single isolates belonging to various CCs.

Proteases. The target sequence of aureolysin, a metalloprotease aur-consensus were found in comparable amounts in the African and German isolates. In African isolates two other proteases, serineproteases *spl*A (66% vs. 56%, $p < 0.0001$) and slpB were significantly more often detected (72% vs. 56%, $p < 0.0001$). In contrast the third serineprotease *spl*E were predominant in German isolates (39% vs. 53%, $p < 0.0001$) and nasal isolates (42% vs. 50%, $p < 0.01$).

set/ssl genes. This group of genes encoding of superantigen/superantigen-like proteins revealed a marked heterogeneity, both with respect to the recognition of the various targets of the same gene as well as of the different genes represented on the Identibac chip. Overall, the differences between African and German isolates in recognition of the entire sets of these genes and alleles were minor, particularly in the interesting group of clinical isolates. Remarkably, with exception of few alleles, CC152 isolates did not yield a positive signal for most of the set/ssl alleles tested.

Capsule genes / biofilm associated genes. The signals recognizing the *cap*5 and *cap*8 target were associated with CCs with CC5, CC8, CC22, and CC152 carrying type 5 capsule genes while the others with exception of CC1 carried type 8 genes. CC1 was found to be inhomogeneous with the majority of isolates (21/27) demonstrating type 8 type, yet a smaller number (6/27) was positive for type 5. In line, the overall presence of these genes in the various geographic or clinical significance groups depends rather on the representation of the CCs: for instance, in commensal samples, the capsule type 5 encoding genes were predominant in isolates from African volunteers (43% vs. 33%, $p < 0.01$) while capsule type 8 encoding genes were more frequent in isolates from Germany (67% vs. 56%, $p < 0.01$). All isolates were positive for *ica*A and *ica*D. DNA from CC152 isolates did not hybridize with the *ica*C target. In contrast, a signal for

the bap gene associated with bovine *S. aureus* strains was absent in all 1200 isolates investigated.

Adhesion factors / MSCRAMM proteins. Similarly to the set/ssl gene cluster, this group of genes revealed a heterogeneous result when probed with the DNA of study isolates. While for some genes a positive signal could be detected in the large majority of isolates (e.g. the consensus targets for bbp, *clf*A, *clf*B, *ebp*, *eno*, *fnb*A, *sdr*C and *Sdr*D, and *vwb*), detection of *fib* (a gene encoding for a fibrinogen binding protein), *ebh* (encoding for the cell wall associated fibronectin-binding protein), *fnb*B (the gene encoding for a fibronectin binding protein), and *sas*G (*Staphylococcus aureus* surface protein G) were ascertained more frequently in African isolates (p¡0.0001). In contrast the map (major histocompatibility complex class II analog protein) or also called *eap* (extracellular adhesion protein) were more often detected in German isolates ($p < 0.0001$). In contrast to the differences observed in the above-mentioned genes or gene alleles (with target recognition more or less frequent throughout several CCs), these overall differences in adhesive proteins have in common that they are caused by a numeric effect of isolates belonging to one CC or few CCs whose DNA is not recognized by this target of interest for array hybridization, yet, with a numerical isolate imbalance within these specific CCs with respect to isolate provenience (Africa/Germany). For instance, the fib gene is not recognized in CC22, yet with 37 German and only 4 African isolates belonging to CC22, this skews the overall result towards a larger fib positivity among African isolates. In line, for *map/eap* (a gene which in previous studies has been shown to be present throughout the *S. aureus* species) the overall 18% *map/eap* negative isolates in the African cohort (as compared to only 3%in the German cohort) are largely caused by the entire lack of *map/eap* target recognition in all isolates of the CC152 complex which, however, is being represented by 83 African and only 2 German isolates. In addition, the reduced portion of *map/eap* positive isolates from Africa in the CC1 complex further contributes to the overall different result. map/eap recognition in the other CCs approached 100% irrespective of isolate origin. It is therefore suggested that in particular within the group of these repeat-rich adhesin genes, gene polymorphisms unique to clones or geographic clades and not represented on the array may contribute to this result. In order to find support for this hypothesis, we examined those exemplar isolates of the CC152 complex which had undergone WGS, and found indeed an aberrant *map/eap* gene contained in their genome (not shown). The *vwb* gene tested revealed no differences between the groups when taking into account all alleles represented on the array.

mprF, isdA, and lmrP genes. These groups of genes also revealed no significant difference when accounting for all alleles tested.

hsd type I restriction enzymes. This class of genes revealed some significant differences of gene allele positivity between isolates of African versus German origin, however, in part they were compensated by recognition of additional allele target. No clear trend or conclusion can be made.

hyaluronate lyse genes. Similarly to the *hsd* type genes, significant differences between isolate groups were found for a given allele, yet, upon consideration of the target recognition over several alleles represented on the array, these differences were compensated.

Association analysis of genes with isolate clusters. Principal component analysis (Figure 6.3) showed the 1200 isolates each characterized by 333 DNA microarray target sequences in a two dimensional space. Isolates were discriminated according to their genetic background (clonal complex) in six separate isolate clusters (3-8), on the

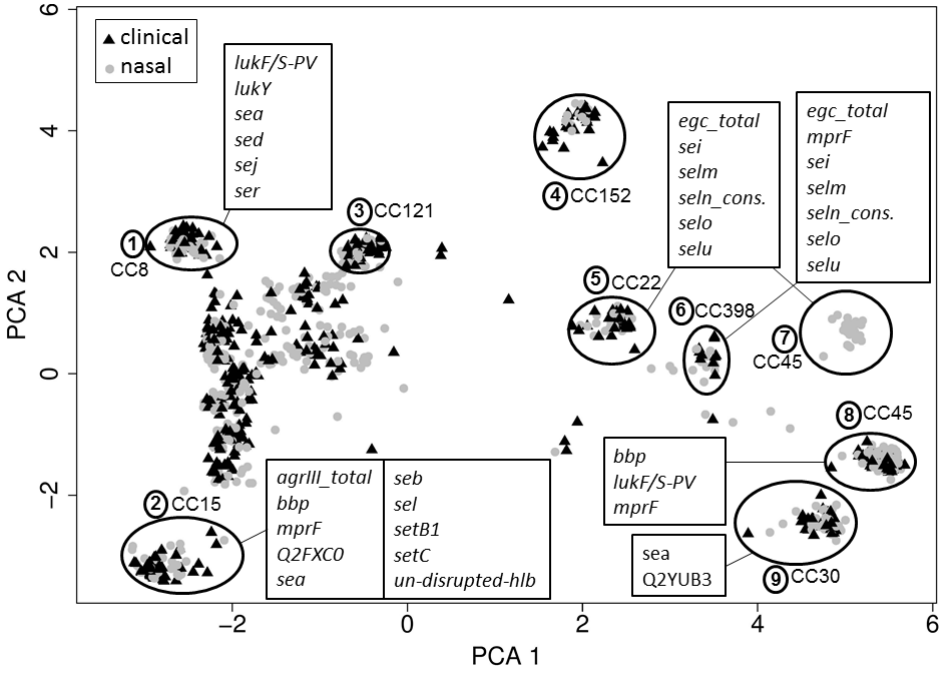


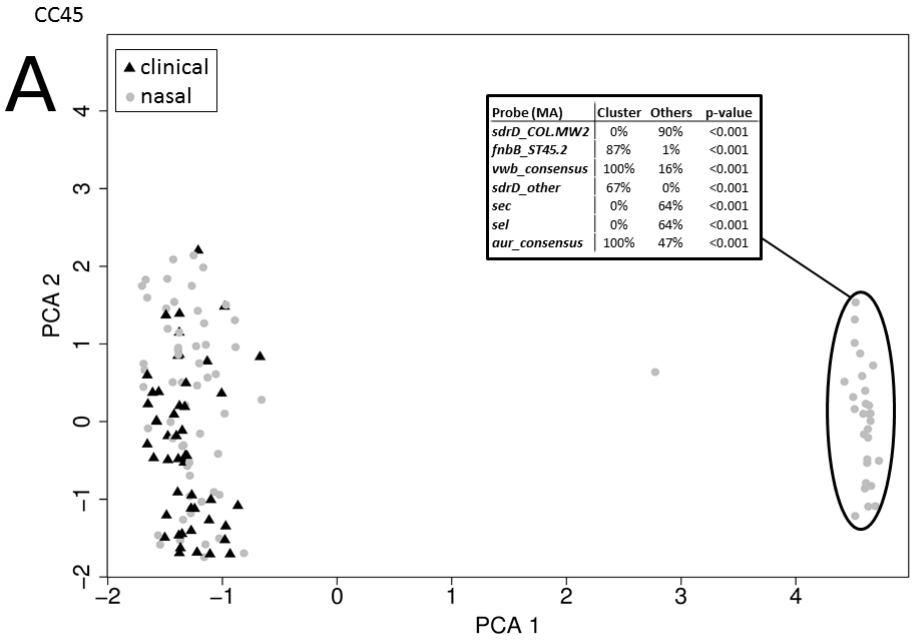
FIGURE 6.3: Two dimensional ordination of 600 clinical and 600 nasal *S. aureus* isolates based on the first two axes of a principal component analysis corresponding to the presence or absence of the 333 target sequences, determined by DNA Microarray. Genes responsible for significant differences ($p < 0.01$) which were found in less then four clusters were shown in boxes.

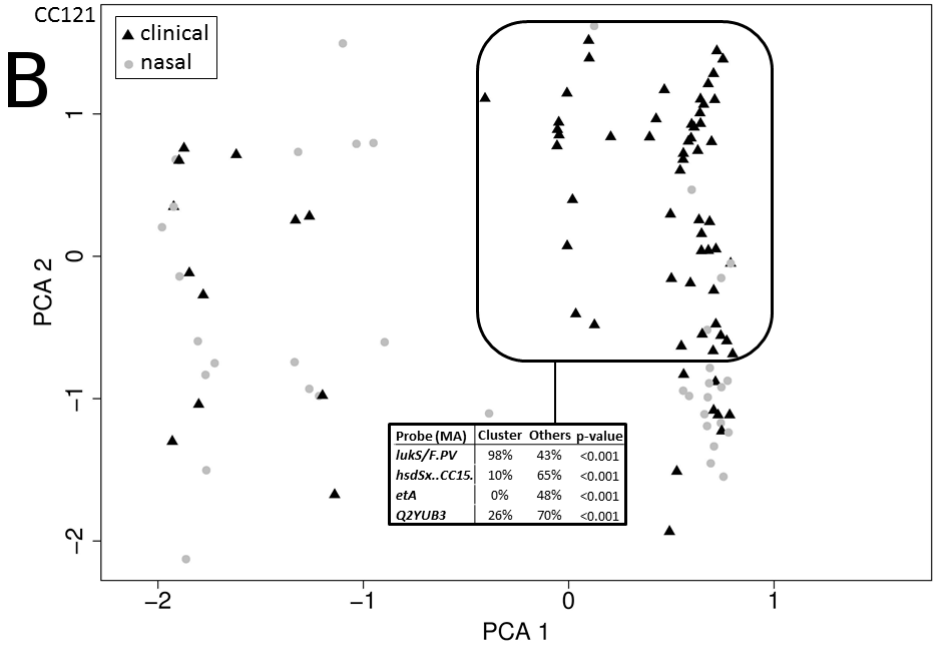
right plot side and two clusters (1, 2) on the left plot side. Further a heterogeneous isolate group belonging to different clonal complexes could be seen on the left plot side without grouping in specific genotypic clusters. The silhouette analysis underlines that each isolate lays well within the defined clusters with a silhouette value of 0.59, minimum and maximum the same.

No separation of isolates according to geographical origin or previous localization in or on the body could be seen, except of 24 CC45 nasal samples which cluster independently of a second heterogeneous CC45 cluster of isolates of different clinical origin. Kolmogorov-Smirnoff test was performed to analyze which genes/ alleles are responsible for significant differences between the eight detected PCA isolate clusters. Cluster 1 (CC8) were specified by the presence of the genes encoding the PVL (*luk*F/S-PV), leucocidin Y (*luk*Y) and the staphylococcal enterotoxins *sea*, *sej*, *ser*. CC15 isolates of cluster 2 carry the genes or target sequences of *agr*III total, *bbp* (bone sialoprotein-binding protein), *fnb*A (fibronectin A), *luk*Y, the target sequence for the hypothetical protein Q2FXC0, staphylococcal exotoxin like protein (*set*B1) and for the undisrupted haemolysin b. CC152 isolates of cluster 3 were the gene *sak* (staphylokinase) while the characteristic genes of CC22 (cluster 4) and of the nasal isolates of the CC45 (cluster 6) were genes of the enterotoxin gene cluster (*egc*) which are sei, selm, seln-cons, selo and selu. Isolates of the zoonotic CC398 of cluster 5 were characterized by the carriage of the resistance operon *bla*I, *bla*R1, *bla*Z encoding for the beta-lactamase resistance the egc target sequence and single genes of the egc cluster, as well as the gene mprF, encoding for mupirocin resistance and sak. The second, heterogenous cluster of CC45 isolates

(cluster 7) were marked by the genes *bla*I, *bla*R1, *bla*Z, *fnb*A, *luk*F/S-PV, *mpr*F. The CC30 isolates of the last cluster 8 were characterized by staphylococcal enterotoxin sea.

Subgroup analysis of clonal complexes associated with clinical isolate origin. The clonal complexes (Figure 6.4) CC45 (Figure 6.4 A), CC101 (Figure 6.4 B), CC121 (Figure 6.4 C) and CC152 (Figure 6.4 D) were found to be either associated to clinical associated isolates or nasal origin. Because of this PCA were performed for isolates of these CCs to identify potential isolate localization specific subgroups and subgroup specific genes. PCA plots show that origin specific clustering of isolates could be seen in CC45 (Figure 6.4 A) and CC121 (Figure 6.4 B) while in the PCAs of CC101 (Figure 6.4 C) and CC152 (Figure 6.4 D) such a clustering could not be found. The nasal CC45 isolates were specified by the presence of target sequences of the von-Willebrandt-factor vwb consensus and aureolysin (aur), the presence of the Fibronectin binding protein B allele fnbB ST45 and of the sdrD other (Ser-Asp rich fibrinogen-/bone sialoprotein-binding protein D) target sequence. Further the absence of the enterotoxin genes sec, sel and the allele sdrD-COL-MW2. The clinical isolate cluster of CC121 were associated with the presence of *luk*F- and *luk*S-PV and characterized by the minor presence of hsdSx-CC15, the exfoliative toxin etA and the target sequence for the hypothetical protein Q2YUB3 according to all other isolates outside of the predominantly clinical cluster.


CC45
A
clinical
nasal
Probe (MA) Cluster Others p-value
sdrD_COL.MW2 0% 90% <0.001
fnbB_ST45.2 87% 1% <0.001
vwb_consensus 100% 16% <0.001
sdrD_other 67% 0% <0.001
sec 0% 64% <0.001
sel 0% 64% <0.001
aur_consensus 100% 47% <0.001
PCA 2
PCA 1


CC121
B
clinical
nasal
Probe (MA) Cluster Others p-value
lukS/F.PV 98% 43% <0.001
hsdSx..CC15. 10% 65% <0.001
etA 0% 48% <0.001
Q2YUB3 26% 70% <0.001
PCA 2
PCA 1

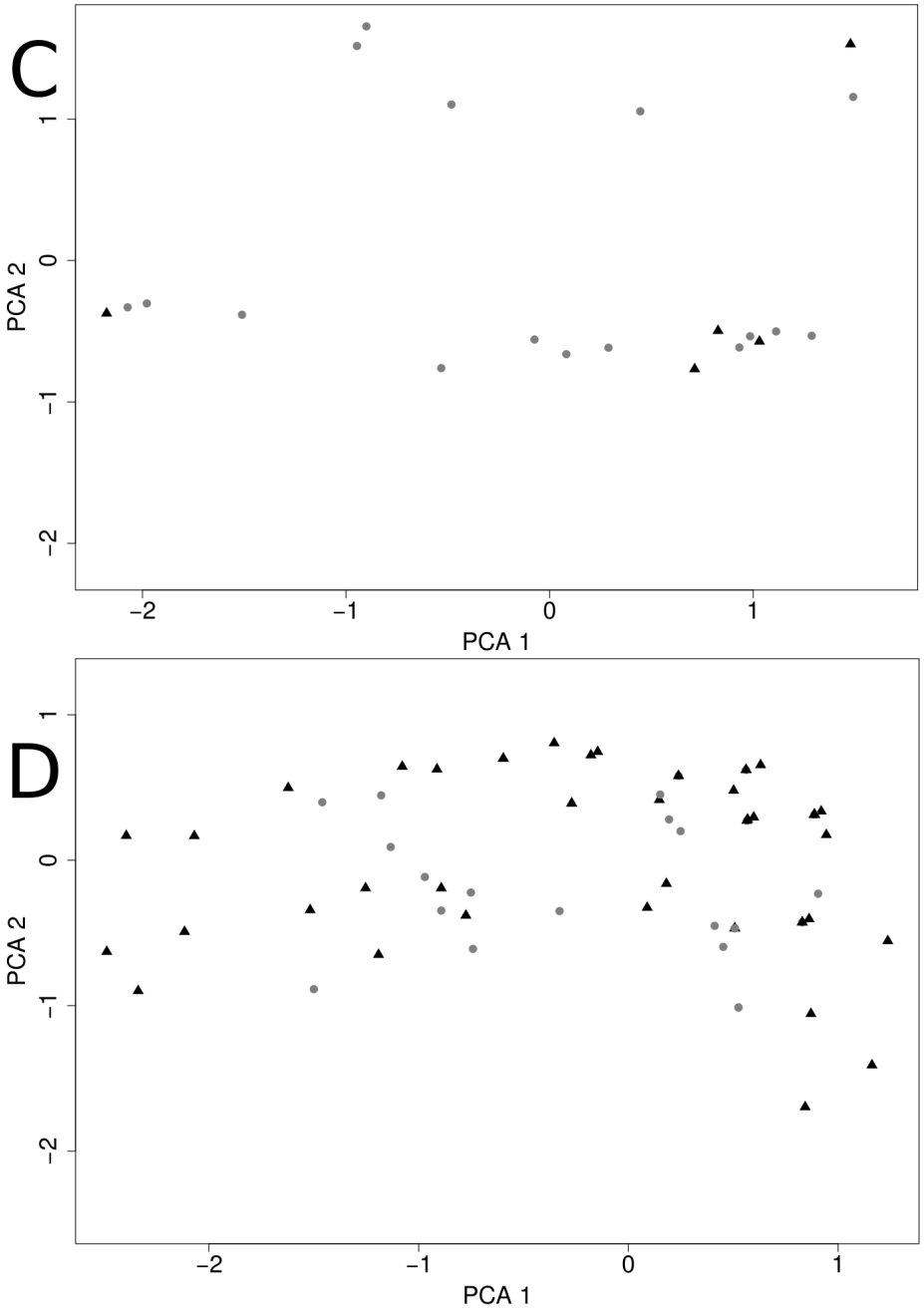


FIGURE 6.4: Principal component analysis based on the first two axes corresponding to the presence or absence of the 333 target sequences which were determined by DNA microarray of the *S. aureus* isolates of A, CC45; B CC121; C CC101 and D CC152. Genes/Alleles written in the boxes are specific for the joined isolate cluster.

6.3 Discussion

Although current studies show that community associated MRSA are on the rise there are still only a limited number of studies available examining the topic of community associated *S. aureus*. Independent of this in developing countries, as in sub-Saharan Africa epidemiological and molecular data of comparable methods of *S. aureus* infections and nasal isolates are rare.

To the best of our knowledge this is the first African-German multicenter study of community associated *S. aureus* isolates of her size determining the geographic differences of the clonal structure of *S. aureus* in three sub-Saharan Africa and German regions with special regard to the original isolate localization site and addressing the question of gene associations to the isolate origin based on one comparable molecular characterization method.

In our study we identified 22 MSSA CCs with at least six isolates in the African and German study sites showing the geographic independent *S. aureus* diversity in the African and German community. Further it has to be stated that the diversity is independent of the previous isolate localization, if they are of clinical or nasal origin. The MRSA isolates in contrast could be only assigned to 8 different CCs showing in accordance to previous suggestions that the acquisition of the SCCmec cassette carrying the methicillin resistance mecA is restricted to special *S. aureus* lineages [65] or that is has to be triggered by antibiotic treatment.

Population studies of *S. aureus* identified five main genotypic clusters CC5, CC8, CC22, CC30 and CC45 as the essential genetic backgrounds of *S. aureus* with differences in the local prevalence in Europe and the USA as well as in Indonesia [152]. In our study these clonal complexes belong to the 10 predominant CCs (Figures 6.1 and 6.5) completed by three other pandemic clones CC1, CC80 and CC121. For all these CCs the dissemination in different African countries had been shown before [21, 147, 185, 140].

The most predominant CC in this study, CC15 was significantly higher prevalent in the African study cohort but with no significant higher proportion of clinical or nasal isolates. This clonal complex is reported worldwide (www.mlst.net) and identified as predominant CC in previous studies of Mali, former Portuguese colonies as Angola, Cape Verde and Sao Tom and the United Kingdom [36, 62, 175]. Moreover CC15 as well as CC30, CC121 and CC152 are known PVL-positive clones as shown in our study, as well [21, 74].

CC22, CC30 and CC45 known as typical MRSA lineages in the Euregio-Meuse-Rhine region [42] were in correspondence to these findings more often found in the German isolates. But interestingly the isolates were mainly Methicillin-susceptible while most available studies described them as Methicillin-resistant strains what could be explained by the fact that these were mainly studies of isolates collected in hospitals [23, 49].

CC121 as well as isolates of CC15 were rarely identified as MRSA [44, 109] according to our study findings. In previous studies it has been investigated why the SCCmec dont integrate in the CC121 genotypes but the answer could not be found.

ST152/CC152 is considered as the major clone of CA-MRSA in the Balkan region and is also responsible for cases of PVL positive CA-MRSA infection in Central Europe and is supposed to originate in Africa, migrated through central Europe and acquired the methicillin resistance [154, 157]. A study from Mali showed that it is the second most frequent MSSA lineage isolated from healthy carriers with 100% PVL positive

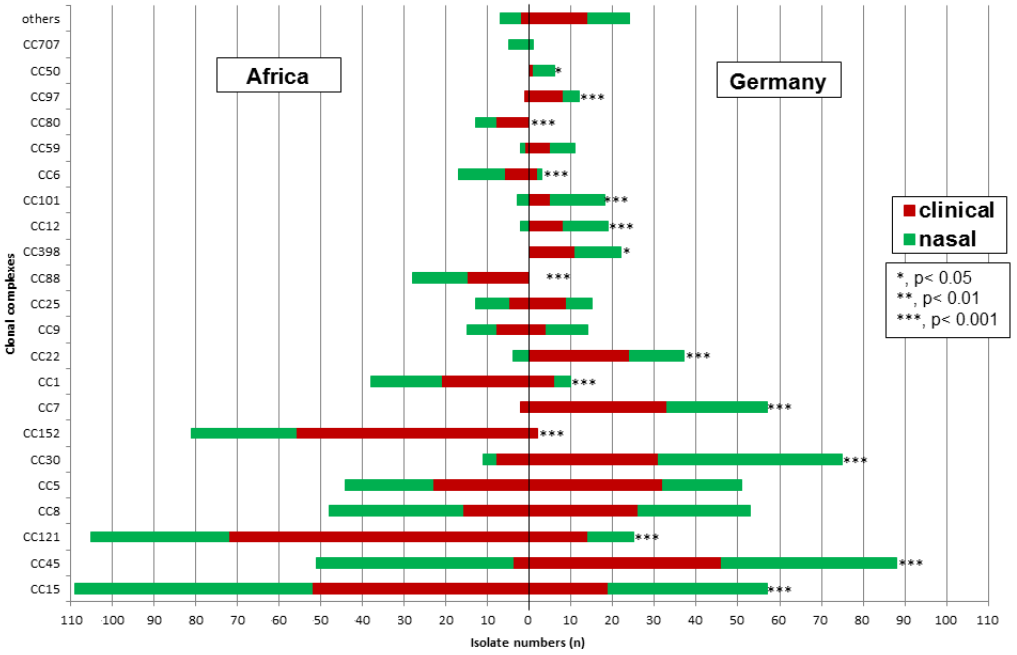


FIGURE 6.5: Distribution of the 22 most prevalent clonal complexes in the African and German study sites. Clonal complexes and sequence types with not at least six isolates where taken together as others. The clonal complexes in the y-axis were ordered according to the total number of identified isolates independent of the study site from the bottom to the top. Prevalence of clinical (red) and nasal (green) isolates in the African (left side) and German group (right side) is shown. Fishers exact test; *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

isolates [175] and taken together the results of different studies it could be considered that CC152 is the major PVL positive clonal complex in sub-Saharan Africa [179]. These is supported by our data which shows that we have only two German CC152 isolates while all other isolates were PVL positive and of African as well as notably of clinical infection sites.

Surprisingly ST80/CC80 known as the most frequently reported community-associated MRSA clone in Europe, also identified in Australia [38], the USA [192] and the UK [133] in this study ST80/CC80 MSSA which are not common [157] has been found in the African cohort as in previous African countries [145, 163] and while strains are not common.

In African studies the MRSA isolates mostly could be assigned to the multi locus sequence type CC88 in South Africa, CC88 in Nigeria as well as in five African towns of a multicenter study [21, 156]. Among all isolates of the study *agr* types (*agr*I to *agr*IV), *agr*IV were identified with an overrepresentation in African and agrI in German isolates. The different agr types are known to be associated with some diseases as e.g. *agr*IV is associated with exfoliatin production [92] and the association of *agr*I and II with reduced vancomycin susceptibility [177]. It has been suggested that the genome of a certain *agr*-group has specific gene combinations that give rise to a specific phenotype [63] or regulate specific gene combinations as described before [22, 52]. In our study a

geographic and maybe host specific factors. The same can be postulated for the both detected capsule types, capsule type 5, predominant in African nasal or capsule type 8 predominant in German nasal samples which were found in all examined isolates of the study.

Comparable results the capsule type distribution for an African cohort and European data overall have been shown in a study of remote African pygmies as well [180]. Both capsule types have been reported that they were the only capsular serotypes associated with human disease [130, 205, 114]. Although the both capsule types could only be detected with geographic dependent predominance in nasal isolates it can be supposed that these were the predominant capsule types of colonizing *S. aureus* strains.

The MRSA prevalence in our study was very low (3%) compared to data of USA, Europe, and previous African studies of Southwestern Nigeria (1.4 to 50%), a meta-analysis in Central Africa (27.7%) and another study on clinical *S. aureus* isolates from this region (11.1%) or in a study on neonatal bloodstream infection on Tanzania (28%) [131, 198, 181, 1, 191]. The reason for the low MRSA rate could be the defined inclusion criteria of the case related forms, that the participants were only included if they didnt have contact with clinical institutions during the last 6 months. This let suggest that they didnt get antibiotics and they should not be in contact with hospital-associated *S. aureus* strains and therefore shouldnt be colonized by them.

In accordance to the phenotypic data of many African studies showing a high resistance to penicillin ($73.7-100\%$) [101, 163, 179] and tetracycline ($21.8-92\%$). [127, 48] found a significant predominance of the beta lactamase operon and of the tetracycline encoding genes *tet*K and *tet*M in the African isolates particularly pronounced in special CCs what has not been described like that before for African isolates. The frequent prescription of aminopenicillins and the use of tetracycline in livestock [54]) maybe could explain the high resistance towards these antimicrobial agents in sub-Sahara Africa. Interestingly the beta lactamase operon was predominant in nasal CC45 but in clinical CC152 African isolates but with a higher relative abundance of the resistance genes in the CC152 isolates. This lead to the assumption that CC45 is a more colonizing strain while CC152 strains in combination with the PVL encoding genes are strains with a higher risk of causing infections, adopted to the possible treatment.

In previous studies [123, 5] *erm*A was described as the dominant erythromycin resistance gene in their investigated MRSA isolates while the prevalence of ermB in *S. aureus* was less than 2%. It has been found that the *erm*A were carried on a transposon (Tn554) [41] while *erm*C is typically located on a plasmid (pUSA03) [45]. In the present study we found a geographic dependent predominance of the erythromycin resistance genes *erm*A in German isolates and of *erm*C in African isolates [125] while *erm*B were not found in contrast to a study of a Algiers hospital and Taiwanese studies investigating CA-MRSA. In these studies they found that *erm*B was more widespread than *erm*A or *erm*C [201, 125, 47], maybe this is caused by a region specific distribution of the mobile genetic elements carrying *erm*A or *erm*C. A review considering the antibiotic susceptibility profile of MRSA in Africa [60] showed that $84-99\%$ of the MRSA isolates were fosfomycin susceptible. For MSSA isolates data of a remote pygmie study showed fosfomycin susceptibility in all isolates as well (Schaumburg, PlosNegTropDis, 2011). In this study the fosfomycin resistance gene *fos*B was detected in even a higher proportion in African isolates especially of clinical origin but in lower numbers in African CC1 than in German CC1 isolates. This might suggest that the fosfomycin sensitive strains could be a reservoir for the development of fosfomycin resistant MRSA and that different CC1

lineages were spread in the African and German regions.

A prominent finding of this study is the high prevalence of the Panton Valentine encoding genes *luk*S/F-PV, identified in 287 mainly African isolates (24%) of the CCs 15, 88, 121 and 152. African field studies of the last two decades showed that Africa is a PVL-endemic region with high rates of PVL in pandemic CA-MRSA- and MSSA-lineages [26, 129, 180, 175, 154] but mainly in MSSA ranging from 17 to 74%. This is not in agreement with other European findings with PVL prevalences between 1 to 5% [133, 199, 126]. The higher rates of PVL positive African isolates in our study group correspond to the previous African studies and significant PVL prevalence differences between developing countries and the German study sites as example for an industrialized country has been seen before [153]. But the question why there is a higher predominance of PVL in Africa is still unanswered. Based on previous studies it has been hypothesized that host factors, such as an altered C5a receptor, unidentified *S. aureus* virulence factors or the humid environment of tropical Africa could be parameters has to be considered [117, 186]. Taken together PVL positive MSSA are a likely reservoir for the development of PVL positive MRSA [165] why surveillance of developing MRSA-PVL positive isolates would be important to control the possible rise.

As well as PVL the leukocidins *Luk*D-*Luk*E are more prevalent in clinical than in carrier isolates and could be isolated from different kinds of staphylococcal infections [15]. In this study in contrast we didnt see such predominance of the corresponding genes *luk*D and *luk*E but for *luk*X and *luk*Y in clinical isolates maybe because we analyzed only community associated isolates. Further a geographic and CC dependent prevalence of *luk*D and *luk*E especially in CC1 could be seen maybe cause by the presence of different lineages in the African and German regions. Overall it could be said that the prevalence of leukocidin genes is higher in African isolates.

The toxic shock syndrome toxin *tst*1 was found in 13 different CCs with an overrepresentation in German isolates ($p < 0.01$) especially in CC30 as described in previous studies [138] that there is an association of *tst*1 and CC30. Although the *tst* gene was less frequently present in the African isolates (6%) than in other investigations of nasal isolates of Ireland, Germany and Poland [35, 13], the data were comparable to a Gabonese study [181]. Interestingly in CC8 the gene was significantly more often found in African isolates what has not described before.

Prevalence reports of pyrogentic toxin superantigens (PTSAg genes) differ depending on the geographic affiliation, the analyzed population structure and the included, tested staphylococcal PTSAg genes. Special enterotoxins were detected as group because they were carried by the same mobile genetic element *sec-sel, sed-sej-ser, sek-seq* and *egc* (*seg, sei, sem, sen, seo, seu*) [86]. Although at least half of the investigated study isolates harbor one of these exotoxins or *sea, seb*. We found geographic and genotypic differences for the PTSAG prevalence in accordance with other studies of different geographic origin [15, 101]. The geographic origin, previous body localization as well as CC specific (e.g. *sea*) enterotoxin abundance differences let assume that the distribution of different mobile genetic elements (MGE) could be responsible for these significant enterotoxin predominances because it could be shown that MGEs carrying PTSAg genes were strongly associated with the clonal background. So *seb* as seen in the pygmies study were predominant in African isolates especially of CC5 and CC121 but further a higher *seb* prevalence in the German isolates in comparison to European data [138] could be found. Overall *seb* has been more often detected in clinical isolates.

Sak, chp, scn and *sea* located on $\beta - hemolysin$converting bacteriophages build an

immune evations cluster (IEC). Different IEC variants were described [197]. It has been shown that each CC lineage has a unique but highly conserved combination of immune evations genes but that they were no differences between MSSA and MRSA or CA-MRSA and HA-MRSA [128] but it has been supposed that the host-pathogen interaction is lineage specific. In this study we found lineage, geographic and previous body localization dependent gene prevalence differences this let suggest that different bacteriophages and immune evasion clusters could be found not only in Africa and Germany but also CC specific in Africa in Germany. Because of this it has to be investigated in more detail the kind and distribution of the bacteriophages with regard to the clinical outcome.

The scalded skin syndrome typically occurring in neonates and infants, but also possible to affect predisposed adults, is caused by the epidermolytic proteases ETA and ETB encoded by the genes *eta* and *etb* [122]. Authors in [159] reported that 22%of invasive strains carried *eta* but not *etb*. The reported prevalences of different studies of clinical isolates were variable according to the study focus [9]. In comparison to literature [15, 86, 145] we have a higher prevalence of *et*A and a similar *et*B rate without geographic specificity but clonal specific overrepresentation in CC15 and CC121 particular pronounced in German CC121 isolates as could be seen before in a study of German healthy carriers [138].

The proteases, as a group, are of great importance to the virulence of the bacterium [214, 167]. Staphylococci are able to secrete up to eight different serine proteases, two cysteine proteases, and one metalloprotease. Two analysed serine proteasse encoded by e.g. *spl*A and *spl*E were significantly more often detected in African isolates while *spl*E were predominant in German and nasal isolates. In a Swedish study comparing commensal and invasive *S. aureus* isolates *spl*A, *spl*B were found to be significantly associated with invasive disease while no significant association with one of the both groups were found for *spl*E (Rasmussen, PlosOne, 2013). This is in contrast to our results, we didnt find such an association but geographic specific differences for the different proteases as we see have not been described till yet based on the fact that only a limited numbers of publications investigating the role of *spl*-genes in infection diseases are available.

We could show that there is a geographic dependent and/or clinical/nasal origin dependent prevalence for specific single genes while the appearance of other genes is strictly in line with the genetic *S. aureus* lineage as has been show by [128, 159]. This goes in line with the global gene prevalence analysis which show that independent of the isolate origin the isolates build genotypic based clusters except of some CC45 isolates. Because of this further expression and phenotypic analysis have to be performed to clarify if there are geographic or nasal/ clinical phenotypic differences and to see who are the responsible factors of the success of CC15, CC121 and CC152 in the African countries while the known pandemic MRSA lineages CC22, CC30 and CC45 arent successful there.

Taken together we can conclude that always the genetic background has to be considered and therefore surveillance taken molecular characteristics into account are of importance for adjusted empirical treatment and to control the distribution of successful newly adapted strains.

Chapter 7

Conclusions and outlook

7.1 BEclear package

We developed a novel tool called BEclear that reduces the negative impact of batch effect on DNA methylation data sets. It is array platform independent. We tested the devised methodology on breast invasive carcinoma data from The Cancer Genome Atlas and compared it with the existing algorithms ComBat, Surrogate Variable Analysis and Functional normalization. BEclear outperformed these methods in terms of precision and avoids changing the unaffected data. BEclear is available as an R package at the Bioconductor project [1]

7.2 AKSmooth: enhancing low-coverage bisulfite sequencing data via kernel-based smoothing

A method of growing importance is called $WGBS$ - Whole-genome bisulfite sequencing. It exclusively provides a consistent outlook on the genome-wide DNA methylation profile. Yet, normally to get a sufficient amount of genome and read coverage it involves high sequencing costs. The efforts of bioinformaticians to postprocess sequencing data and thereby increase its quality can lower this costs. Thus, our method called Adjusted Local Kernel Smoother or AKSmooth is aimed to embody this. It is a statistical approach which is able to reconstruct the single CpG methylation estimate across the entire methylome using low-coverage bisulfite sequencing ($Bi-Seq$) data consistently and efficiently. We have showed its performance on the low-coverage ($\sim 4\times$) DNA methylation profiles of three human colon cancer samples and matched controls [31].

Having used diverse parameters, we received high concordance with the gold standard high-coverage sample (Pearson 0.90), outperforming the popular analogous method BSmooth (for AKSmooth-curated). AKSmooth reported computational efficiency with runtime benchmark over 4.5 times better than the reference tool.

AKSmooth turned out to be a simple and resultant tool that can provide an accurate human colon methylome estimation profile from low-coverage WGBS data. Moreover, its implementation is available in R package [2].

[1] http://bioconductor.org/packages/release/bioc/html/BEclear.html
[2] https://github.com/Junfang/AKSmooth

7.3 Outlook

Several future research projects aim to exploit the practical utility of results highlighted in current thesis, in terms of both batch effect correction and *S. aureus* data analysis. These projects intend to study deeper mechanisms of cancerogenesis as well as applying other data mining tools to *S. aureus* data, namely redescription mining and classification trees, in order to connect clinical data with the bacterial genome.

On the one hand, the performance of BEclear can still be improved by optimizing its programming code. On the other hand, BEclear was designed to be platform independent, that means that it has potential of batch effect adjustment of DNA methylation data obtained from other technologies than Infinium HumanMethylation450 array. Of special interest would be the application of BEclear to the Whole genome bisulfite sequencing data since it is considered as the gold standard in producing high resolution epigenomic data. Furthermore, batch effect detection and correction in other epigenetic changes, such as chromatin modifications, is still not completely investigated. Thus I believe, that additional methods developed within this PhD project and now available as software packages will contribute for progress in all areas of epigenetic research.

Bibliography

[1] S. Adesida, H. Boelens, B. Babajide, A. Kehinde, S. Snijders, W. V. Leeuwen, A. Coker, H. Verbrugh, and A. V. Belkum. Major epidemic clones of *Staphylococcus aureus* in Nigeria. *Microbial Drug Resistance*, 11(2):115–121, 2005.

[2] A. Ako-Nai, A. Ogunniyi, A. Lamikanra, and S. Torimiro. The characterisation of clinical isolates of *Staphylococcus aureus* in Ile-Ife, Nigeria. *Journal of Medical Microbiology*, 34(2):109–112, 1991.

[3] R. Akulenko and V. Helms. DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. *Human Molecular Genetics*, 22(15):3016–3022, 2013.

[4] P. G. Alluri, I. A. Asangani, and A. M. Chinnaiyan. BETs abet Tam-R in ER - positive breast cancer. *Cell Research*, 24(8):899–900, 2014.

[5] L. S. Almer, V. D. Shortridge, A. M. Nilius, J. M. Beyer, N. B. Soni, M. H. Bui, G. G. Stone, and R. K. Flamm. Antimicrobial susceptibility and molecular characterization of community-acquired methicillin-resistant *Staphylococcus aureus*. *Diagnostic Microbiology and Infectious Disease*, 43(3):225–232, 2002.

[6] J. Anguzu and D. Olila. Drug sensitivity patterns of bacterial isolates from septic postoperative wounds in a regional referral hospital in Uganda. *African Health Sciences*, 7(3), 2007.

[7] A. C. Antoniou, O. M. Sinilnikova, L. McGuffog, S. Healey, H. Nevanlinna, T. Heikkinen, J. Simard, A. B. Spurdle, J. Beesley, X. Chen, et al. Common variants in LSP1, 2q35 and 8q24 and breast cancer risk for BRCA1 and BRCA2 mutation carriers. *Human Molecular Genetics*, 18(22):4442–4456, 2009.

[8] S. L. Anzick, J. Kononen, R. L. Walker, D. O. Azorsa, M. M. Tanner, X.-Y. Guan, G. Sauter, O.-P. Kallioniemi, J. M. Trent, and P. S. Meltzer. AIB1, a steroid receptor coactivator amplified in breast and ovarian cancer. *Science*, 277(5328):965–968, 1997.

[9] M. Argudín, M. C. Mendoza, F. Méndez, M. C. Martín, B. Guerra, and M. R. Rodicio. Clonal complexes and diversity of exotoxin gene profiles in methicillin-resistant and methicillin-susceptible *Staphylococcus aureus* isolates from patients in a Spanish hospital. *Journal of Clinical Microbiology*, 47(7):2097–2105, 2009.

[10] M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen, and R. A. Irizarry. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014.

[11] N. Ashraf, S. Zino, A. Macintyre, D. Kingsmore, A. Payne, W. George, and P. Shiels. Altered sirtuin expression is associated with node-positive breast cancer. *British Journal of Cancer*, 95(8):1056–1061, 2006.

[12] F. J. Azuaje, H. Wang, H. Zheng, F. Léonard, M. Rolland-Turner, L. Zhang, Y. Devaux, and D. R. Wagner. Predictive integration of gene functional similarity and co-expression defines treatment response of endothelial progenitor cells. *BMC Systems Biology*, 5(1):46, 2011.

[13] J. Bania, A. Dabrowska, K. Korzekwa, A. Zarczynska, J. Bystron, J. Chrzanowska, and J. Molenda. The profiles of enterotoxin genes in *Staphylococcus aureus* from nasal carriers. *Letters in Applied Microbiology*, 42(4):315–320, 2006.

[14] Z. Barekati, R. Radpour, C. Kohler, B. Zhang, P. Toniolo, P. Lenner, Q. Lv, H. Zheng, and X. Y. Zhong. Methylation profile of TP53 regulatory pathway and mtDNA alterations in breast cancer patients lacking TP53 mutations. *Human Molecular gGnetics*, 19(15):2936–2946, 2010.

[15] K. Becker, A. W. Friedrich, G. Lubritz, M. Weilert, G. Peters, and C. Von Eiff. Prevalence of genes encoding pyrogenic toxin superantigens and exfoliative toxins among strains of *Staphylococcus aureus* isolated from blood and nasal specimens. *Journal of Clinical Microbiology*, 41(4):1434–1439, 2003.

[16] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[17] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, et al. High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–295, 2011.

[18] A. Bird. DNA methylation patterns and epigenetic memory. *Genes & Development*, 16(1):6–21, 2002.

[19] Z. Birnbaum, F. H. Tingey, et al. One-sided confidence contours for probability distribution functions. *The Annals of Mathematical Statistics*, 22(4):592–596, 1951.

[20] C. Bock and T. Lengauer. Computational epigenetics. *Bioinformatics*, 24(1):1–10, 2008.

[21] S. Breurec, S. Zriouil, C. Fall, P. Boisier, S. Brisse, S. Djibo, J. Etienne, M. Fonkoua, J. Perrier-Gros-Claude, R. Pouillot, et al. Epidemiology of methicillin-resistant *Staphylococcus aureus* lineages in five major African towns: emergence and spread of atypical clones. *Clinical Microbiology and Infection*, 17(2):160–165, 2011.

[22] S. Bronner, H. Monteil, and G. Prévost. Regulation of virulence determinants in *Staphylococcus aureus*: complexity and applications. *FEMS Microbiology Reviews*, 28(2):183–200, 2004.

[23] A. Budimir, R. Deurenberg, Z. Bošnjak, E. Stobberingh, H. Cetkovic, and S. Kalenic. A variant of the Southern German clone of methicillin-resistant *Staphylococcus aureus* is predominant in Croatia. *Clinical Microbiology and Infection*, 16(8):1077–1083, 2010.

[24] D. V. Bulavin, O. N. Demidov, S. Saito, P. Kauraniemi, C. Phillips, S. A. Amundson, C. Ambrosino, G. Sauter, A. R. Nebreda, C. W. Anderson, et al. Amplification of PPM1D in human tumors abrogates p53 tumor-suppressor activity. *Nature Genetics*, 31(2):210–215, 2002.

[25] P. Bushel. pvca: Principal Variance Component Analysis (PVCA). R package version 1.10.0, 2013.

[26] S. J. Campbell, H. S. Deshmukh, C. L. Nelson, I.-G. Bae, M. E. Stryjewski, J. J. Federspiel, G. T. Tonthat, T. H. Rude, S. L. Barriere, R. Corey, et al. Genotypic characteristics of *Staphylococcus aureus* isolates from a multinational trial of complicated skin and skin structure infections. *Journal of Clinical Microbiology*, 46(2):678–684, 2008.

[27] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[28] A. Catteau and J. R. Morris. BRCA1 methylation: a significant role in tumour development? In *Seminars in Cancer Biology*, volume 12, pages 359–371. Elsevier, 2002.

[29] M. Chan, G. Liang, and P. Jones. Relationship between transcription and DNA methylation. *Current Topics in Microbiology and Immunology*, 249:75–86, 2000.

[30] S. J. Chanock, L. Burdett, M. Yeager, V. Llaca, A. Langerød, S. Presswalla, R. Kaaresen, R. L. Strausberg, D. S. Gerhard, V. Kristensen, et al. Somatic sequence alterations in twenty-one genes selected by expression profile analysis of breast carcinomas. *Breast Cancer Research*, 9(1):R5, 2007.

[31] J. Chen, P. Lutsik, R. Akulenko, J. Walter, and V. Helms. AKSmooth: Enhancing low-coverage bisulfite sequencing data via kernel-based smoothing. *Journal of Bioinformatics and Computational Biology*, 12(06):1442005, 2014. PMID: 25553811.

[32] R. Z. Chen, U. Pettersson, C. Beard, L. Jackson-Grusby, and R. Jaenisch. DNA hypomethylation leads to elevated mutation rates. *Nature*, 395(6697):89–93, 1998.

[33] K. Chheng, S. Tarquinio, V. Wuthiekanun, L. Sin, J. Thaipadungpanit, P. Amornchai, N. Chanpheaktra, S. Tumapa, H. Putchhat, N. Day, et al. Emergence of community-associated methicillin-resistant *Staphylococcus aureus* associated with pediatric infection in cambodia. *PLoS One*, 4(8):e6630, 2009.

[34] Y.-Y. Chuang and Y.-C. Huang. Molecular epidemiology of community-associated meticillin-resistant *Staphylococcus aureus* in Asia. *The Lancet Infectious Diseases*, 13(8):698–708, 2013.

[35] M. M. Collery, D. S. Smyth, J. M. Twohig, A. C. Shore, D. C. Coleman, and C. J. Smyth. Molecular typing of nasal carriage isolates of *Staphylococcus aureus* from an Irish university student population based on toxin gene PCR, agr locus types and multiple locus, variable number tandem repeat analysis. *Journal of Medical Microbiology*, 57(3):348–358, 2008.

[36] T. Conceição, C. Coelho, I. S. Silva, H. de Lencastre, and M. Aires-de Sousa. *Staphylococcus aureus* in Portuguese former colonies from Africa and the far East: missing data to help fill the world map. *Clinical Microbiology and Infection*, 21(9):842, 2015.

[37] W. Conover. *Practical nonparametric statistics*. Wiley, 1971.

[38] G. W. Coombs, G. R. Nimmo, J. M. Bell, F. Huygens, F. G. O'brien, M. J. Malkowski, J. C. Pearson, A. J. Stephens, P. M. Giffard, et al. Genetic diversity among community methicillin-resistant *Staphylococcus aureus* strains causing outpatient infections in Australia. *Journal of Clinical Microbiology*, 42(10):4735–4743, 2004.

[39] J. Cui, A. C. Antoniou, G. S. Dite, M. C. Southey, D. J. Venter, D. F. Easton, G. G. Giles, M. R. McCredie, and J. L. Hopper. After BRCA1 and BRCA2 - what next? Multifactorial segregation analyses of three-generation, population-based Australian families affected by female breast cancer. *The American Journal of Human Genetics*, 68(2):420–431, 2001.

[40] J. H. Degnan and N. A. Rosenberg. Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2(5):e68, 2006.

[41] R. Deurenberg, C. Vink, S. Kalenic, A. Friedrich, C. Bruggeman, and E. Stobberingh. The molecular evolution of methicillin-resistant *Staphylococcus aureus*. *Clinical Microbiology and Infection*, 13(3):222–235, 2007.

[42] R. H. Deurenberg, E. Nulens, H. Valvatne, S. Sebastian, C. Driessen, J. Craeghs, E. De Brauwer, B. Heising, Y. J. Kraat, J. Riebe, et al. Cross-border dissemination of methicillin-resistant *Staphylococcus aureus*, Euregio Meuse-Rhin region. *Emerging Infectious Diseases*, 15(5):727, 2009.

[43] R. H. Deurenberg, M. I. Rijnders, S. Sebastian, M. A. Welling, P. S. Beisser, and E. E. Stobberingh. The *Staphylococcus aureus* lineage-specific markers collagen adhesin and toxic shock syndrome toxin 1 distinguish multilocus sequence typing clonal complexes within spa clonal complexes. *Diagnostic Microbiology and Infectious Disease*, 65(2):116–122, 2009.

[44] R. H. Deurenberg and E. E. Stobberingh. The evolution of *Staphylococcus aureus*. *Infection, Genetics and Evolution*, 8(6):747–763, 2008.

[45] B. A. Diep, S. R. Gill, R. F. Chang, T. H. Phan, J. H. Chen, M. G. Davidson, F. Lin, J. Lin, H. A. Carleton, E. F. Mongodin, et al. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus*. *The Lancet*, 367(9512):731–739, 2006.

[46] B. A. Diep, G. G. Stone, L. Basuino, C. J. Graber, A. Miller, S.-A. des Etages, A. Jones, A. M. Palazzolo-Ballance, F. Perdreau-Remington, G. F. Sensabaugh, et al. The arginine catabolic mobile element and staphylococcal chromosomal cassette mec linkage: convergence of virulence and resistance in the USA300 clone of methicillin-resistant *Staphylococcus aureus*. *Journal of Infectious Diseases*, 197(11):1523–1530, 2008.

[47] N. Djahmi, N. Messad, S. Nedjai, A. Moussaoui, D. Mazouz, J.-L. Richard, A. Sotto, and J.-P. Lavigne. Molecular epidemiology of *Staphylococcus aureus* strains isolated from inpatients with infected diabetic foot ulcers in an Algerian university hospital. *Clinical Microbiology and Infection*, 19(9):E398–E404, 2013.

[48] F. Djoudi, C. Bonura, S. Benallaoua, A. Touati, D. Touati, A. Aleo, C. Cala, T. Fasciana, and C. Mammina. Panton-Valentine leukocidin positive sequence type 80 methicillin-resistant *Staphylococcus aureus* carrying a staphylococcal cassette chromosome mec type IVc is dominant in neonates and children in an Algiers hospital. *New Microbiologica*, 36(1):49–55, 2013.

[49] G. Donker, R. Deurenberg, C. Driessen, S. Sebastian, S. Nys, and E. Stobberingh. The population structure of *Staphylococcus aureus* among general practice patients from the Netherlands. *Clinical Microbiology and Infection*, 15(2):137–143, 2009.

[50] R. Dreos, G. Ambrosini, R. C. Périer, and P. Bucher. The Eukaryotic promoter database: expansion of EPDnew and new promoter analysis tools. *Nucleic Acids Research*, 43(D1):D92–D96, 2015.

[51] P. Dunman, W. Mounts, F. McAleese, F. Immermann, D. Macapagal, E. Marsilio, L. McDougal, F. Tenover, P. Bradford, P. Petersen, et al. Uses of *Staphylococcus aureus* GeneChips in genotyping and genetic composition analysis. *Journal of Clinical Microbiology*, 42(9):4275–4283, 2004.

[52] P. á. Dunman, E. Murphy, S. Haney, D. Palacios, G. Tucker-Kellogg, S. Wu, E. Brown, R. Zagursky, D. Shlaes, and S. Projan. Transcription profiling-based identification of *Staphylococcus aureus* genes regulated by the agr and/or sarA loci. *Journal of Bacteriology*, 183(24):7341–7353, 2001.

[53] J. Durbin. *Distribution theory for tests based on the sample distribution function*, volume 9. Siam, 1973.

[54] B. Egyir, L. Guardabassi, J. Esson, S. S. Nielsen, M. J. Newman, K. K. Addo, and A. R. Larsen. Insights into nasal carriage of *Staphylococcus aureus* in an urban and a rural community in Ghana. *PloS One*, 9(4), 2014.

[55] M. Ehrlich, M. A. Gama-Sosa, L.-H. Huang, R. M. Midgett, K. C. Kuo, R. A. McCune, and C. Gehrke. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Research*, 10(8):2709–2721, 1982.

[56] E. Eisenberg and E. Y. Levanon. Human housekeeping genes, revisited. *Trends in Genetics*, 29(10):569–574, 2013.

[57] S. Enany, E. Yaoita, Y. Yoshida, M. Enany, and T. Yamamoto. Molecular characterization of Panton-Valentine leukocidin-positive community-acquired methicillin-resistant *Staphylococcus aureus* isolates in Egypt. *Microbiological Research*, 165(2):152–162, 2010.

[58] M. C. Enright, N. P. Day, C. E. Davies, S. J. Peacock, and B. G. Spratt. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. *Journal of Clinical Microbiology*, 38(3):1008–1015, 2000.

[59] M. Esteller, M. Sanchez-Cespedes, R. Rosell, D. Sidransky, S. B. Baylin, and J. G. Herman. Detection of aberrant promoter hypermethylation of tumor suppressor genes in serum DNA from non-small cell lung cancer patients. *Cancer Research*, 59(1):67–70, 1999.

[60] M. E. Falagas, I. P. Korbila, A. Kapaskelis, K. Manousou, L. Leontiou, and G. S. Tansarli. Trends of mortality due to septicemia in greece: an 8-year analysis. *PLoS One*, 8(7):e67621, 2013.

[61] J. G. Falls, D. J. Pulford, A. A. Wylie, and R. L. Jirtle. Genomic imprinting: implications for human disease. *The American Journal of Pathology*, 154(3):635–647, 1999.

[62] E. J. Feil, J. E. Cooper, H. Grundmann, D. A. Robinson, M. C. Enright, T. Berendt, S. J. Peacock, J. M. Smith, M. Murphy, B. G. Spratt, et al. How clonal is *Staphylococcus aureus*? *Journal of Bacteriology*, 185(11):3307–3316, 2003.

[63] Y. Feng, F. Zheng, X. Pan, W. Sun, C. Wang, Y. Dong, A.-p. Ju, J. Ge, D. Liu, C. Liu, et al. Existence and characterization of allelic variants of sao, a newly identified surface protein from *Streptococcus suis*. *FEMS Microbiology Letters*, 275(1):80–88, 2007.

[64] J.-P. Fortin, A. Labbe, M. Lemire, B. W. Zanke, T. J. Hudson, E. J. Fertig, C. M. Greenwood, and K. D. Hansen. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology*, 15(12):503, 2014.

[65] P. Francois, S. Harbarth, A. Huyghe, G. Renzi, M. Bento, A. Gervaix, D. Pittet, and J. Schrenzel. Methicillin-resistant *Staphylococcus aureus*, Geneva, Switzerland, 1993–2005. *Emerging Infectious Diseases*, 14(2):304, 2008.

[66] P. Francois, A. Huyghe, Y. Charbonnier, M. Bento, S. Herzig, I. Topolski, B. Fleury, D. Lew, P. Vaudaux, S. Harbarth, et al. Use of an automated multiple-locus, variable-number tandem repeat-based method for rapid and high-throughput genotyping of *Staphylococcus aureus* isolates. *Journal of Clinical Microbiology*, 43(7):3346–3355, 2005.

[67] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[68] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

[69] B. Ghebremedhin, M. Olugbosi, A. Raji, F. Layer, R. Bakare, B. König, and W. König. Emergence of a community-associated methicillin-resistant *Staphylococcus aureus* strain with a unique resistance profile in Southwest Nigeria. *Journal of Clinical Microbiology*, 47(9):2975–2980, 2009.

[70] H. Golmoghaddam, A. M. Pezeshki, A. Ghaderi, and M. Doroudchi. CD1a and CD1d genes polymorphisms in breast, colorectal and lung cancers. *Pathology & Oncology Research*, 17(3):669–675, 2011.

[71] G. Golub and W. Kahan. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial & Applied Mathematics, Series B: Numerical Analysis*, 2(2):205–224, 1965.

[72] R. J. Gorwitz. A review of community-associated methicillin-resistant *Staphylococcus aureus* skin and soft tissue infections. *The Pediatric Infectious Disease Journal*, 27(1):1–7, 2008.

[73] R. Goud, S. Gupta, U. Neogi, D. Agarwal, K. Naidu, R. Chalannavar, and G. Subhaschandra. Community prevalence of methicillin and vancomycin resistant *Staphylococcus aureus* in and around Bangalore, southern india. *Revista da Sociedade Brasileira de Medicina Tropical*, 44(3):309–312, 2011.

[74] H. Grundmann, D. M. Aanensen, C. C. Van Den Wijngaard, B. G. Spratt, D. Harmsen, A. W. Friedrich, E. S. R. L. W. Group, et al. Geographic distribution of *Staphylococcus aureus* causing invasive infections in Europe: a molecular-epidemiological analysis. *PLoS Medicine*, 7(1):49, 2010.

[75] D. J. Gubler. Current research on dengue. In *Current Topics in Vector Research*, pages 37–56. Springer, 1987.

[76] Z. Guo, J. Yuan, W. Tang, X. Chen, X. Gu, K. Luo, Y. Wang, B. Wan, and L. Yu. Cloning and characterization of the human gene RAP2C, a novel member of ras family, which activates transcriptional activities of sre. *Molecular Biology Reports*, 34(3):137–144, 2007.

[77] C. J. Hackbarth and H. F. Chambers. blaI and blaR1 regulate beta-lactamase and PBP 2a production in methicillin-resistant *Staphylococcus aureus*. *Antimicrobial Agents and Chemotherapy*, 37(5):1144–1149, 1993.

[78] A. Hamdan-Partida, T. Sainz-Espuñes, and J. Bustos-Martínez. Characterization and persistence of *Staphylococcus aureus* strains isolated from the anterior nares and throats of healthy carriers in a Mexican community. *Journal of Clinical Microbiology*, 48(5):1701–1705, 2010.

[79] D. Harmsen, H. Claus, W. Witte, J. Rothgänger, H. Claus, D. Turnwald, and U. Vogel. Typing of methicillin-resistant *Staphylococcus aureus* in a university hospital setting by using novel software for spa repeat determination and database management. *Journal of Clinical Microbiology*, 41(12):5442–5448, 2003.

[80] S. Harris, E. Feil, M. Holden, M. Quail, E. Nickerson, N. Chantratita, S. Gardete, A. Tavares, N. Day, J. Lindsay, J. Edgeworth, H. de Lencastre, J. Parkhill, S. Peacock, and S. Bentley. Evolution of mrsa during hospital transmission and intercontinental 338 spread. *Science*, 327(5964):469–474, 2010.

[81] N. V. Hayes, E. Blackburn, L. V. Smart, M. M. Boyle, G. A. Russell, T. M. Frost, B. J. Morgan, A. J. Baines, and W. J. Gullick. Identification and characterization of novel spliced variants of neuregulin 4 in prostate cancer. *Clinical Cancer Research*, 13(11):3147–3155, 2007.

[82] J. G. Herman, A. Umar, K. Polyak, J. R. Graff, N. Ahuja, J.-P. J. Issa, S. Markowitz, J. K. Willson, S. R. Hamilton, K. W. Kinzler, et al. Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proceedings of the National Academy of Sciences USA*, 95(12):6870–6875, 1998.

[83] M. J. Hoenerhoff, I. Chu, D. Barkan, Z. Liu, S. Datta, G. P. Dimri, and J. E. Green. BMI1 cooperates with H-RAS to induce an aggressive breast cancer phenotype with brain metastases. *Oncogene*, 28(34):3022–3032, 2009.

[84] K. Holm, C. Hegardt, J. Staaf, J. Vallon-Christersson, G. Jönsson, H. Olsson, Å. Borg, and M. Ringnér. Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. *Breast Cancer Research*, 13(3):R36, 2010.

[85] A. Holmes, G. Edwards, E. Girvan, W. Hannant, J. Danial, J. Fitzgerald, and K. Templeton. Comparison of two multilocus variable-number tandem-repeat methods and pulsed-field gel electrophoresis for differentiating highly clonal methicillin-resistant *Staphylococcus aureus* isolates. *Journal of Clinical Microbiology*, 48(10):3600–3607, 2010.

[86] S. Holtfreter, D. Grumann, M. Schmudde, H. Nguyen, P. Eichler, B. Strommenger, K. Kopron, J. Kolata, S. Giedrys-Kalemba, I. Steinmetz, et al. Clonal distribution of superantigen genes in clinical *Staphylococcus aureus* isolates. *Journal of Clinical Microbiology*, 45(8):2669–2680, 2007.

[87] K. Horan, C. Jang, J. Bailey-Serres, R. Mittler, C. Shelton, J. F. Harper, J.-K. Zhu, J. C. Cushman, M. Gollery, and T. Girke. Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiology*, 147(1):41–57, 2008.

[88] Y.-H. Hsu, C.-H. Hsing, C.-F. Li, C.-H. Chan, M.-C. Chang, J.-J. Yan, and M.-S. Chang. Anti–IL-20 monoclonal antibody suppresses breast cancer progression and bone osteolysis in murine models. *The Journal of Immunology*, 188(4):1981–1991, 2012.

[89] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *Eighth IEEE International Conference on Data Mining ICDM'08*, pages 263–272. IEEE, 2008.

[90] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2008.

[91] D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267, 2006.

[92] S. Jarraud, G. Lyon, A. Figueiredo, L. Gérard, F. Vandenesch, J. Etienne, T. Muir, and R. Novick. Exfoliatin-producing strains define a fourthagr specificity group in *Staphylococcus aureus*. *Journal of Bacteriology*, 182(22):6517–6522, 2000.

[93] S. Jarraud, C. Mougel, J. Thioulouse, G. Lina, H. Meugnier, F. Forey, X. Nesme, J. Etienne, and F. Vandenesch. Relationships between *Staphylococcus aureus* genetic background, virulence factors, agr groups (alleles), and human disease. *Infection and Immunity*, 70(2):631–641, 2002.

[94] W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[95] P. A. Jones and S. B. Baylin. The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics*, 3(6):415–428, 2002.

[96] M. Kawaguchiya, N. Urushibara, O. Kuwahara, M. Ito, K. Mise, and N. Kobayashi. Molecular characteristics of community-acquired methicillin-resistant *Staphylococcus aureus* in Hokkaido, northern main island of Japan: Identification of sequence types 6 and 59 Panton-Valentine Leucocidin–positive community-acquired methicillin-resistant *Staphylococcus aureus*. *Microbial Drug Resistance*, 17(2):241–250, 2011.

[97] M. J. Kelley, S. Li, and D. H. Harpole. Genetic analysis of the β-tubulin gene, TUBB, in non-small-cell lung cancer. *Journal of the National Cancer Institute*, 93(24):1886–1888, 2001.

[98] C. Klein, G. Georges, K.-P. Künkele, R. Huber, R. A. Engh, and S. Hansen. High thermostability and lack of cooperative DNA binding distinguish the p63 core domain from the homologous tumor suppressor p53. *Journal of Biological Chemistry*, 276(40):37390–37401, 2001.

[99] R. Köck, K. Becker, B. Cookson, J. van Gemert-Pijnen, S. Harbarth, J. Kluytmans, M. Mielke, G. Peters, R. Skov, M. Struelens, et al. Methicillin-resistant *Staphylococcus aureus* (MRSA): burden of disease and control challenges in Europe. *Eurosurveillance*, 15(41), 2010.

[100] R. Köck, K. Siam, S. Al-Malat, J. Christmann, F. Schaumburg, K. Becker, and A. Friedrich. Characteristics of hospital patients colonized with livestock-associated meticillin-resistant *Staphylococcus aureus* (MRSA) CC398 versus other MRSA clones. *Journal of Hospital Infection*, 79(4):292–296, 2011.

[101] D. O. Kolawole, A. Adeyanju, F. Schaumburg, A. L. Akinyoola, O. O. Lawal, Y. B. Amusa, R. Kock, and K. Becker. Characterization of colonizing *Staphylococcus aureus* isolated from surgical wards patients in a Nigerian university hospital. *PLoS One*, 8(7):e68721, 2013.

[102] L. Koreen, S. V. Ramaswamy, E. A. Graviss, S. Naidich, J. M. Musser, and B. N. Kreiswirth. spa typing method for discriminating among *Staphylococcus aureus* isolates: implications for use of a single marker to detect genetic micro-and macrovariation. *Journal of Clinical Microbiology*, 42(2):792–799, 2004.

[103] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[104] C. U. Köser, M. T. Holden, M. J. Ellington, E. J. Cartwright, N. M. Brown, A. L. Ogilvy-Stuart, L. Y. Hsu, C. Chewapreecha, N. J. Croucher, S. R. Harris, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *New England Journal of Medicine*, 366(24):2267–2275, 2012.

[105] S. Kriaucionis and A. Bird. DNA methylation and Rett syndrome. *Human Molecular Genetics*, 12(suppl 2):R221–R227, 2003.

[106] M. Krupp, T. Maass, J. U. Marquardt, F. Staib, T. Bauer, R. König, S. Biesterfeld, P. R. Galle, A. Tresch, and A. Teufel. The functional cancer map: a systems-level synopsis of genetic deregulation in cancer. *BMC Medical Genomics*, 4(1):53, 2011.

[107] N. B. Kuemmerle, E. Rysman, P. S. Lombardo, A. J. Flanagan, B. C. Lipe, W. A. Wells, J. R. Pettus, H. M. Froehlich, V. A. Memoli, P. M. Morganelli, et al. Lipoprotein lipase links dietary fat to solid tumor cell proliferation. *Molecular Cancer Therapeutics*, 10(3):427–436, 2011.

[108] M. Kulis and M. Esteller. DNA methylation and cancer. *Advances in Genetics*, 70:27–56, 2010.

[109] K. Kurt, J.-P. Rasigade, F. Laurent, R. V. Goering, H. Zemlickova, I. Machova, M. J. Struelens, A. E. Zautner, S. Holtfreter, B. Broker, et al. Subpopulations of *Staphylococcus aureus* clonal complex 121 are associated with distinct clinical entities. *PLoS One*, 8(3):e58155, 2013.

[110] G. La Rocca, R. Anzalone, S. Corrao, F. Magno, F. Rappa, S. Marasa, A. Czarnecka, L. Marasa, C. Sergi, G. Zummo, et al. CD1a down-regulation in primary invasive ductal breast carcinoma may predict regional lymph node invasion and patient outcome. *Histopathology*, 52(2):203–212, 2008.

[111] P. W. Laird. Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11(3):191–203, 2010.

[112] B. Lamy, F. Laurent, O. Gallon, F. Doucet-Populaire, J. Etienne, J.-W. Decousser, C. de Bactériologie Virologie Hygiène (ColBVH) Study Group, et al. Antibacterial resistance, genes encoding toxins and genetic background among *Staphylococcus aureus* isolated from community-acquired skin and soft tissue infections in France: a national prospective survey. *European Journal of Clinical Microbiology & Infectious Diseases*, 31(6):1279–1284, 2012.

[113] G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies II. Clustering systems. *The Computer Journal*, 10(3):271–277, 1967.

[114] S. M. Lattar, L. P. Tuchscherr, R. L. Caccuri, D. Centrón, K. Becker, C. A. Alonso, C. Barberis, G. Miranda, F. R. Buzzola, C. von Eiff, et al. Capsule expression and genotypic differences among *Staphylococcus aureus* isolates from patients with chronic or acute osteomyelitis. *Infection and Immunity*, 77(5):1968–1975, 2009.

[115] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey. Package sva, 2014.

[116] J. Li and R. Tibshirani. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research*, 22(5):519–536, 2013.

[117] M. Li, X. Du, A. E. Villaruz, B. A. Diep, D. Wang, Y. Song, Y. Tian, J. Hu, F. Yu, Y. Lu, et al. MRSA epidemic linked to a quickly spreading colonization and virulence determinant. *Nature Medicine*, 18(5):816–819, 2012.

[118] Y. Li, J. Zhu, G. Tian, N. Li, Q. Li, M. Ye, H. Zheng, J. Yu, H. Wu, J. Sun, et al. The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biology*, 8(11):e1000533, 2010.

[119] H. M. Lightfoot, A. Lark, C. A. Livasy, D. T. Moore, D. Cowan, L. Dressler, R. J. Craven, and W. G. Cance. Upregulation of focal adhesion kinase (FAK) expression in ductal carcinoma in situ (DCIS) is an early event in breast tumorigenesis. *Breast Cancer Research and Treatment*, 88(2):109–116, 2004.

[120] J. Y. Lim, S. O. Yoon, S. W. Hong, J. W. Kim, S. H. Choi, and J. Y. Cho. Thioredoxin and thioredoxin-interacting protein as prognostic markers for gastric cancer recurrence. *World Journal of Gastroenterology: WJG*, 18(39):5581, 2012.

[121] D. Lin. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304, 1998.

[122] G. Lina, Y. Gillet, F. Vandenesch, M. E. Jones, D. Floret, and J. Etienne. Toxin involvement in staphylococcal scalded skin syndrome. *Clinical Infectious Diseases*, 25(6):1369–1373, 1997.

[123] G. Lina, A. Quaglia, M.-E. Reverdy, R. Leclercq, F. Vandenesch, and J. Etienne. Distribution of genes encoding resistance to macrolides, lincosamides, and streptogramins among staphylococci. *Antimicrobial Agents and Chemotherapy*, 43(5):1062–1066, 1999.

[124] J. A. Lindsay, C. E. Moore, N. P. Day, S. J. Peacock, A. A. Witney, R. A. Stabler, S. E. Husain, P. D. Butcher, and J. Hinds. Microarrays reveal that each of the ten dominant lineages of *Staphylococcus aureus* has a unique combination of surface-associated and regulatory genes. *Journal of Bacteriology*, 188(2):669–676, 2006.

[125] W.-T. Lo, W.-J. Lin, M.-H. Tseng, J.-J. Lu, S.-Y. Lee, M.-L. Chu, and C.-C. Wang. Nasal carriage of a single clone of community-acquired methicillin-resistant *Staphylococcus aureus* among kindergarten attendees in northern Taiwan. *BMC Infectious Diseases*, 7(1):51, 2007.

[126] F. D. Lowy. *Staphylococcus aureus* infections. *New England Journal of Medicine*, 339(8):520–532, 1998.

[127] B. J.-J. Mariem, T. Ito, M. Zhang, J. Jin, S. Li, B.-B. B. Ilhem, H. Adnan, X. Han, and K. Hiramatsu. Molecular characterization of methicillin-resistant Panton-valentine leukocidin positive *Staphylococcus aureus* clones disseminating in Tunisian hospitals and in the community. *BMC Microbiology*, 13(1):2, 2013.

[128] A. J. McCarthy and J. A. Lindsay. *Staphylococcus aureus* innate immune evasion is lineage-specific: a bioinfomatics study. *Infection, Genetics and Evolution*, 19:7–14, 2013.

[129] B. T. Meeren, P. S. Millard, M. Scacchetti, M. H. Hermans, M. Hilbink, T. B. Concelho, J. J. Ferro, and P. C. Wever. Emergence of methicillin resistance and Panton-Valentine leukocidin positivity in hospital-and community-acquired *Staphylococcus aureus* infections in Beira, Mozambique. *Tropical Medicine & International Health*, 19(2):169–176, 2014.

[130] D. C. Melles, K. L. Taylor, A. I. Fattom, and A. Van Belkum. Serotyping of Dutch *Staphylococcus aureus* strains from carriage and infection. *FEMS Immunology & Medical Microbiology*, 52(2):287–292, 2008.

[131] T. V. Mhada, F. Fredrick, M. I. Matee, and A. Massawe. Neonatal sepsis at Muhimbili national hospital, Dar es Salaam, Tanzania; aetiology, antimicrobial sensitivity pattern and clinical outcome. *BMC Public Health*, 12(1):904, 2012.

[132] C. Milheiriço, D. C. Oliveira, and H. de Lencastre. Update to the multiplex PCR strategy for assignment of mec element types in *Staphylococcus aureus*. *Antimicrobial Agents and Chemotherapy*, 51(9):3374–3377, 2007.

[133] S. Monecke, B. Berger-Bächi, G. Coombs, A. Holmes, I. Kay, A. Kearns, H.-J. Linde, F. O'Brien, P. Slickers, and R. Ehricht. Comparative genomics and DNA array-based genotyping of pandemic *Staphylococcus aureus* strains encoding Panton-Valentine leukocidin. *Clinical Microbiology and Infection*, 13(3):236–249, 2007.

[134] S. Monecke, G. Coombs, A. C. Shore, D. C. Coleman, P. Akpaka, M. Borg, H. Chow, M. Ip, L. Jatzwauk, D. Jonas, et al. A field guide to pandemic, epidemic and sporadic clones of methicillin-resistant *Staphylococcus aureus*. *PloS One*, 6(4):e17936, 2011.

[135] S. Monecke and R. Ehricht. Rapid genotyping of methicillin-resistant *Staphylococcus aureus* (MRSA) isolates using miniaturised oligonucleotide arrays. *Clinical Microbiology and Infection*, 11(10):825–833, 2005.

[136] S. Monecke, R. Ehricht, P. Slickers, N. Wiese, and D. Jonas. Intra-strain variability of methicillin-resistant *Staphylococcus aureus* strains ST228-MRSA-I and ST5-MRSA-II. *European Journal of Clinical Microbiology & Infectious Diseases*, 28(11):1383–1390, 2009.

[137] S. Monecke, L. Jatzwauk, S. Weber, P. Slickers, and R. Ehricht. DNA microarray-based genotyping of methicillin-resistant *Staphylococcus aureus* strains from Eastern Saxony. *Clinical Microbiology and Infection*, 14(6):534–545, 2008.

[138] S. Monecke, C. Luedicke, P. Slickers, and R. Ehricht. Molecular epidemiology of *Staphylococcus aureus* in asymptomatic carriers. *European Journal of Clinical Microbiology & Infectious Diseases*, 28(9):1159–1165, 2009.

[139] S. Monecke, P. Slickers, and R. Ehricht. Assignment of *Staphylococcus aureus* isolates to clonal complexes based on microarray analysis and pattern recognition. *FEMS Immunology & Medical Microbiology*, 53(2):237–251, 2008.

[140] A. Moodley, W. Oosthuysen, A. Dusé, E. Marais, et al. Molecular characterization of clinical methicillin-resistant *Staphylococcus aureus* isolates in South Africa. *Journal of Clinical Microbiology*, 48(12):4608–4611, 2010.

[141] P. Moore and J. Lindsay. Genetic variation among hospital isolates of methicillin-sensitive *Staphylococcus aureus*: evidence for horizontal transfer of virulence genes. *Journal of Clinical Microbiology*, 39(8):2760–2767, 2001.

[142] E. Mosca, R. Alfieri, I. Merelli, F. Viti, A. Calabria, and L. Milanesi. A multilevel data integration resource for breast cancer study. *BMC Systems Biology*, 4(1):76, 2010.

[143] D. Nagakubo, T. Taira, H. Kitaura, M. Ikeda, K. Tamai, S. M. Iguchi-Ariga, and H. Ariga. DJ-1, a novel oncogene which transforms mouse NIH3T3 cells in cooperation withras. *Biochemical and Biophysical Research Communications*, 231(2):509–513, 1997.

[144] R. Nantanda, H. Hildenwall, S. Peterson, D. Kaddu-Mulindwa, I. Kalyesubula, and J. K. Tumwine. Bacterial aetiology and outcome in children with severe pneumonia in Uganda. *Annals of Tropical Paediatrics: International Child Health*, 28(4):253–260, 2008.

[145] M. B. Nejma, M. Mastouri, S. Frih, N. Sakly, Y. B. Salem, and M. Nour. Molecular characterization of methicillin-resistant *Staphylococcus aureus* isolated in Tunisia. *Diagnostic Microbiology and Infectious Disease*, 55(1):21–26, 2006.

[146] C. G. A. Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[147] U. A. Ngoa, F. Schaumburg, A. A. Adegnika, K. Kösters, T. Möller, J. F. Fernandes, A. Alabi, S. Issifou, K. Becker, M. P. Grobusch, et al. Epidemiology and population structure of *Staphylococcus aureus* in various population groups from a rural and semi urban area in Gabon, Central Africa. *Acta Tropica*, 124(1):42–47, 2012.

[148] E. K. Nickerson, T. E. West, N. P. Day, and S. J. Peacock. *Staphylococcus aureus* disease and drug resistance in resource-limited countries in south and east Asia. *The Lancet Infectious Diseases*, 9(2):130–135, 2009.

[149] J. Novembre and M. Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5):646–649, 2008.

[150] U. Nübel, A. Nitsche, F. Layer, B. Strommenger, and W. Witte. Single-nucleotide polymorphism genotyping identifies a locally endemic clone of methicillin-resistant *Staphylococcus aureus*. *PLoS One*, 7(3):e32698, 2012.

[151] S. Obaro, L. Lawson, U. Essen, K. Ibrahim, K. Brooks, A. Otuneye, D. Shetima, P. Ahmed, T. Ajose, M. Olugbile, et al. Community acquired bacteremia in young children from central Nigeria-a pilot study. *BMC Infectious Diseases*, 11(1):137, 2011.

[152] F. O'Brien, G. Coombs, J. Pearman, M. Gracey, F. Moss, K. Christiansen, and W. Grubb. Population dynamics of methicillin-susceptible and-resistant *Staphylococcus aureus* in remote communities. *Journal of Antimicrobial Chemotherapy*, page dkp285, 2009.

[153] F. P. O'Hara, N. Guex, J. M. Word, L. A. Miller, J. A. Becker, S. L. Walsh, N. E. Scangarella, J. M. West, R. M. Shawar, and H. Amrine-Madsen. A geographic variant of the staphylococcus aureus Panton-Valentine leukocidin toxin and the origin of community-associated methicillin-resistant *S. aureus* USA300. *Journal of Infectious Diseases*, 197(2):187–194, 2008.

[154] K. O. Okon, P. Basset, A. Uba, J. Lin, B. Oyawoye, A. O. Shittu, and D. S. Blanc. Cooccurrence of predominant Panton-Valentine leukocidin-positive sequence type (ST) 152 and multidrug-resistant ST 241 *Staphylococcus aureus* clones in Nigerian hospitals. *Journal of Clinical Microbiology*, 47(9):3000–3003, 2009.

[155] M. H. Oktay, K. Oktay, D. Hamele-Bena, A. Buyuk, and L. G. Koss. Focal adhesion kinase as a marker of malignant phenotype in breast and cervical carcinomas. *Human Pathology*, 34(3):240–245, 2003.

[156] W. Oosthuysen, H. Orth, C. Lombard, B. Sinha, and E. Wasserman. Population structure analyses of *Staphylococcus aureus* at Tygerberg hospital, South Africa, reveals a diverse population, a high prevalence of Panton–Valentine leukocidin genes, and unique local methicillin-resistant *S. aureus* clones. *Clinical Microbiology and Infection*, 20(7):652–659, 2014.

[157] J. A. Otter and G. L. French. Molecular epidemiology of community-associated meticillin-resistant *Staphylococcus aureus* in Europe. *The Lancet Infectious Diseases*, 10(4):227–239, 2010.

[158] H. S. Parker and J. T. Leek. The practical effect of batch on genomic prediction. *Statistical Applications in Genetics and Molecular Biology*, 11(3):Article 10, 2012.

[159] S. J. Peacock, C. E. Moore, A. Justice, M. Kantzanou, L. Story, K. Mackie, G. O'Neill, and N. P. Day. Virulent combinations of adhesin and toxin genes in natural populations of *Staphylococcus aureus*. *Infection and Immunity*, 70(9):4987–4996, 2002.

[160] I. P. Pogribny, A. G. Basnakian, B. J. Miller, N. G. Lopatina, L. A. Poirier, and S. J. James. Breaks in genomic DNA and within the p53 gene are associated with hypomethylation in livers of folate/methyl-deficient rats. *Cancer Research*, 55(9):1894–1901, 1995.

[161] J. A. Pryer, R. Nichols, P. Elliott, B. Thakrar, E. Brunner, and M. Marmot. Dietary patterns among a national random sample of British adults. *Journal of Epidemiology and Community Health*, 55(1):29–37, 2001.

[162] J. Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6):418–427, 2001.

[163] N. Ramdani-Bouguessa, M. Bes, H. Meugnier, F. Forey, M.-E. Reverdy, G. Lina, F. Vandenesch, M. Tazir, and J. Etienne. Detection of methicillin-resistant *Staphylococcus aureus* strains resistant to multiple antibiotics and carrying the Panton-Valentine leukocidin genes in an Algiers hospital. *Antimicrobial Agents and Chemotherapy*, 50(3):1083–1085, 2006.

[164] R. G. Ramsay and T. J. Gonda. MYB function in normal and cancer cells. *Nature Reviews Cancer*, 8(7):523–534, 2008.

[165] J.-P. Rasigade, F. Laurent, G. Lina, H. Meugnier, M. Bes, F. Vandenesch, J. Etienne, and A. Tristan. Global distribution and evolution of Panton-Valentine leukocidin-positive methicillin-susceptible *Staphylococcus aureus*, 1981–2007. *Journal of Infectious Diseases*, 201(10):1589–1597, 2010.

[166] E. A. Reddy, A. V. Shaw, and J. A. Crump. Community-acquired bloodstream infections in Africa: a systematic review and meta-analysis. *The Lancet Infectious Diseases*, 10(6):417–432, 2010.

[167] S. B. Reed, C. A. Wesson, L. E. Liou, W. R. Trumble, P. M. Schlievert, G. A. Bohach, and K. W. Bayles. Molecular characterization of a novel *Staphylococcus aureus* Serine Protease Operon. *Infection and Immunity*, 69(3):1521–1527, 2001.

[168] W. Reik and J. Walter. Imprinting mechanisms in mammals. *Current Opinion in Genetics & Development*, 8(2):154–164, 1998.

[169] P. Resnik et al. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Ressearch (JAIR)*, 11:95–130, 1999.

[170] B. D. Ripley. *Modern applied statistics with S.* Springer, 2002.

[171] K. D. Robertson and P. A. Jones. DNA methylation: past, present and future directions. *Carcinogenesis*, 21(3):461–467, 2000.

[172] K. D. Robertson, E. Uzvolgyi, G. Liang, C. Talmadge, J. Sumegi, F. A. Gonzales, and P. A. Jones. The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mrna expression in normal tissues and overexpression in tumors. *Nucleic Acids Research*, 27(11):2291–2298, 1999.

[173] D. A. Robinson, A. B. Monk, J. E. Cooper, E. J. Feil, and M. C. Enright. Evolutionary genetics of the accessory gene regulator (agr) locus in *Staphylococcus aureus*. *Journal of Bacteriology*, 187(24):8312–8321, 2005.

[174] U. Ruffing, R. Akulenko, M. Bischoff, V. Helms, M. Herrmann, and L. von Müller. Matched-cohort DNA microarray diversity analysis of methicillin sensitive and methicillin resistant *Staphylococcus aureus* isolates from hospital admission patients. *PLoS One*, 7(12):e2487, 2012.

[175] R. Ruimy, A. Maiga, L. Armand-Lefevre, I. Maiga, A. Diallo, A. K. Koumaré, K. Ouattara, S. Soumaré, K. Gaillard, J.-C. Lucet, et al. The carriage population of *Staphylococcus aureus* from Mali is composed of a combination of pandemic clones and the divergent Panton-Valentine leukocidin-positive genotype ST152. *Journal of Bacteriology*, 190(11):3962–3968, 2008.

[176] V. E. Russo, R. A. Martienssen, A. D. Riggs, et al. *Epigenetic mechanisms of gene regulation*, volume 32. Cold Spring Harbor Laboratory Press, 1996.

[177] G. Sakoulas, G. M. Eliopoulos, R. C. Moellering, C. Wennersten, L. Venkataraman, R. P. Novick, and H. S. Gold. Accessory gene regulator (agr) locus in geographically diverse *Staphylococcus aureus* isolates with reduced susceptibility to vancomycin. *Antimicrobial Agents and Chemotherapy*, 46(5):1492–1502, 2002.

[178] O. Sakwinska, G. Kuhn, C. Balmelli, P. Francioli, M. Giddey, V. Perreten, A. Riesen, F. Zysset, D. S. Blanc, and P. Moreillon. Genetic diversity and ecological success of *Staphylococcus aureus* strains colonizing humans. *Applied and Environmental Microbiology*, 75(1):175–183, 2009.

[179] F. Schaumburg, A. Alabi, G. Peters, and K. Becker. New epidemiology of *Staphylococcus aureus* infection in Africa. *Clinical Microbiology and Infection*, 20(7):589–596, 2014.

[180] F. Schaumburg, R. Köck, A. W. Friedrich, S. Soulanoudjingar, U. A. Ngoa, C. von Eiff, S. Issifou, P. G. Kremsner, M. Herrmann, G. Peters, et al. Population structure of *Staphylococcus aureus* from remote African Babongo Pygmies. *PLoS Neglected Tropical Diseases*, 5(5):e1150–e1150, 2011.

[181] F. Schaumburg, U. A. Ngoa, K. Kösters, R. Köck, A. Adegnika, P. Kremsner, B. Lell, G. Peters, A. Mellmann, and K. Becker. Virulence factors and genotypes of *Staphylococcus aureus* from infection and carriage in Gabon. *Clinical Microbiology and Infection*, 17(10):1507–1513, 2011.

[182] A. Schlicker and M. Albrecht. FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Research*, 36(suppl 1):D434–D439, 2008.

[183] A. Schlicker, F. S. Domingues, J. Rahnenführer, and T. Lengauer. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7(1):302, 2006.

[184] L. M. Schouls, E. C. Spalburg, M. van Luit, X. W. Huijsdens, G. N. Pluister, M. G. van Santen-Verheuvel, H. Van Der Heide, H. Grundmann, M. Heck, and A. J. de Neeling. Multiple-locus variable number tandem repeat analysis of *Staphylococcus aureus*: comparison with pulsed-field gel electrophoresis and spa-typing. *PLoS One*, 4(4):e5082, 2009.

[185] A. O. Shittu, K. Okon, S. Adesida, O. Oyedara, W. Witte, B. Strommenger, F. Layer, and U. Nübel. Antibiotic resistance and molecular epidemiology of *Staphylococcus aureus* in Nigeria. *BMC Microbiology*, 11(1):92, 2011.

[186] A. N. Spaan, T. Henry, W. J. Van Rooijen, M. Perret, C. Badiou, P. C. Aerts, J. Kemmink, C. J. De Haas, K. P. Van Kessel, F. Vandenesch, et al. The staphylococcal toxin Panton-Valentine leukocidin targets human C5a receptors. *Cell Host & Microbe*, 13(5):584–594, 2013.

[187] F. J. Steemers, W. Chang, G. Lee, D. L. Barker, R. Shen, and K. L. Gunderson. Whole-genome genotyping with the single-base extension assay. *Nature Methods*, 3(1):31–33, 2006.

[188] B. Strommenger, C. Braulke, D. Heuck, C. Schmidt, B. Pasemann, U. Nübel, and W. Witte. spa typing of *Staphylococcus aureus* as a frontline tool in epidemiological typing. *Journal of Clinical Microbiology*, 46(2):574–581, 2008.

[189] B. Strommenger, C. Kettlitz, T. Weniger, D. Harmsen, A. Friedrich, and W. Witte. Assignment of Staphylococcus isolates to groups by spa typing, SmaI macrorestriction analysis, and multilocus sequence typing. *Journal of Clinical Microbiology*, 44(7):2533–2540, 2006.

[190] Z. Sun, H. S. Chai, Y. Wu, W. M. White, K. V. Donkena, C. J. Klein, V. D. Garovic, T. M. Therneau, and J.-P. A. Kocher. Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC Medical Genomics*, 4(1):84, 2011.

[191] S. Taiwo, M. Bamidele, E. Omonigbehin, K. Akinsinde, S. Smith, B. Onile, and A. Olowe. Molecular epidemiology of methicillan-resistant *Staphylococcus aureus* in Ilorin, Nigeria. *West African Journal of Medicine*, 24(2):100–106, 2005.

[192] F. C. Tenover, L. K. McDougal, R. V. Goering, G. Killgore, S. J. Projan, J. B. Patel, and P. M. Dunman. Characterization of a strain of community-associated methicillin-resistant *Staphylococcus aureus* widely disseminated in the United States. *Journal of Clinical Microbiology*, 44(1):108–118, 2006.

[193] M. Tessema, C. M. Yingling, M. J. Grimes, C. L. Thomas, Y. Liu, S. Leng, N. Joste, and S. A. Belinsky. Differential epigenetic regulation of TOX subfamily high mobility group box genes in lung and breast cancers. *PloS One*, 7(4):e34850, 2012.

[194] M. Toyota and J.-P. J. Issa. CpG island methylator phenotypes in aging and cancer. In *Seminars in Cancer Biology*, volume 9, pages 349–357. Elsevier, 1999.

[195] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA*, 98(9):5116–5121, 2001.

[196] J. Udo, M. Anah, S. Ochigbo, I. Etuck, and A. Ekanem. Neonatal morbidity and mortality in Calabar, Nigeria: a hospital-based study. *Nigerian Journal of Clinical Practice*, 11(3):285–289, 2008.

[197] W. J. van Wamel, S. H. Rooijakkers, M. Ruyken, K. P. van Kessel, and J. A. van Strijp. The innate immune modulators staphylococcal complement inhibitor and chemotaxis inhibitory protein of *Staphylococcus aureus* are located on β-hemolysin-converting bacteriophages. *Journal of Bacteriology*, 188(4):1310–1315, 2006.

[198] E. Vlieghe, M. Phoba, J. M. Tamfun, and J. Jacobs. Antibiotic resistance among bacterial pathogens in Central Africa: a review of the published literature between 1955 and 2008. *International Journal of Antimicrobial Agents*, 34(4):295–303, 2009.

[199] C. von Eiff, A. W. Friedrich, G. Peters, and K. Becker. Prevalence of genes encoding for members of the staphylococcal leukotoxin family among clinical isolates of *Staphylococcus aureus*. *Diagnostic Microbiology and Infectious Disease*, 49(3):157–162, 2004.

[200] H. Vu-Thien, K. Hormigos, G. Corbineau, B. Fauroux, H. Corvol, D. Moissenet, G. Vergnaud, and C. Pourcel. Longitudinal survey of *Staphylococcus aureus* in cystic fibrosis patients using a multiple-locus variable-number of tandem-repeats analysis method. *BMC Microbiology*, 10(1):24, 2010.

[201] C.-C. Wang, W.-T. Lo, M.-L. Chu, and L. Siu. Epidemiological typing of community-acquired methicillin-resistant *Staphylococcus aureus* isolates from children in Taiwan. *Clinical Infectious Diseases*, 39(4):481–487, 2004.

[202] J. Wang, W. W. Tsang, and G. Marsaglia. Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, 8(18), 2003.

[203] X. Wang, R. C. Southard, C. D. Allred, D. R. Talbert, M. E. Wilson, and M. W. Kilgore. MAZ drives tumor-specific expression of PPAR gamma 1 in breast cancer cells. *Breast Cancer Research and Treatment*, 111(1):103–111, 2008.

[204] D. Watson, R. Elton, W. Jack, J. Dixon, U. Chetty, and W. Miller. The H-ras oncogene product p21 and prognosis in human breast cancer. *Breast Cancer Research and Treatment*, 17(3):161–169, 1991.

[205] A. Watts, D. Ke, Q. Wang, A. Pillay, A. Nicholson-Weller, and J. C. Lee. *Staphylococcus aureus* strains that express serotype 5 or serotype 8 capsular polysaccharides differ in virulence. *Infection and Immunity*, 73(6):3502–3511, 2005.

[206] C. Wi. *Practical nonparametric statistics*. New York, Wiley, 1971.

[207] X. Xiao, D. Li, L. Gao, X. Li, Q. Wang, S. Zhang, and Z. Liu. Screening for cancer associated MiRNAs through co-gene, co-function and co-pathway analysis. *Computers in Biology and Medicine*, 42(5):624–630, 2012.

[208] Y. Xu, T. A. Bismar, J. Su, B. Xu, G. Kristiansen, Z. Varga, L. Teng, D. E. Ingber, A. Mammoto, R. Kumar, et al. Filamin A regulates focal adhesion disassembly and suppresses breast cancer cell migration and invasion. *The Journal of Experimental Medicine*, 207(11):2421–2437, 2010.

[209] T. Xue, Y. You, D. Hong, H. Sun, and B. Sun. The *Staphylococcus aureus* KdpDE two-component system couples extracellular K+ sensing and Agr signaling to infection programming. *Infection and Immunity*, 79(6):2154–2167, 2011.

[210] W. Yang, K. Yoshigoe, X. Qin, J. S. Liu, J. Y. Yang, A. Niemierko, Y. Deng, Y. Liu, A. K. Dunker, Z. Chen, et al. Identification of genes and pathways involved in kidney renal clear cell carcinoma. *BMC Bioinformatics*, 15(Suppl 17):S2, 2014.

[211] X. Yang, L. Yan, and N. Davidson. DNA methylation in breast cancer. *Endocrine-Related Cancer*, 8(2):115–127, 2001.

[212] J. A. Yoder, C. P. Walsh, and T. H. Bestor. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, 13(8):335–340, 1997.

[213] B. C. Young, T. Golubchik, E. M. Batty, R. Fung, H. Larner-Svensson, A. A. Votintseva, R. R. Miller, H. Godwin, K. Knox, R. G. Everitt, et al. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proceedings of the National Academy of Sciences USA*, 109(12):4550–4555, 2012.

[214] M. Zdzalik, A. Y. Karim, K. Wolski, P. Buda, K. Wojcik, S. Brueggemann, P. Wojciechowski, S. Eick, A.-M. Calander, I.-M. Jonsson, et al. Prevalence of genes encoding extracellular proteases in *Staphylococcus aureus*–important targets triggering immune response in vivo. *FEMS Immunology & Medical Microbiology*, 66(2):220–229, 2012.

*

Appendix A

Supplementary material of chapter 6

spa-type	*spa*-type repeat succession	CC	SCC*mec*-type	*agr*-type	toxin profile	resistance profile	strain assignment	Microarray - cluster dendrogram
t003	26-17-20-17-12-17-17-16	CC5	II	II	sed, sej, ser, egc	blaZ/I/R, ermA, aadD, tetEfflux, fosB	Rhine-Hesse	R22
t003	26-17-20-17-12-17-17-16	CC5	II	II	sea, sed, sej, ser egc	blaZ/I/R, ermA, aadD, tetEfflux, fosB	Rhine-Hesse	R24
t003	26-17-20-17-12-17-17-16	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, ermA, aadD, tetEfflux, fosB	Rhine-Hesse	R23
t003	26-17-20-17-12-17-17-16	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, aacA-aphD, aadD, ermA, tetEfflux, cat, fosB	Rhine-Hesse	R15
t003	26-17-20-17-12-17-17-16	CC5	II	II	sed, sej, ser, egc	blaZ/I/R, ermA, aadD, tetEfflux, fosB	Rhine-Hesse	R16
t003	26-17-20-17-12-17-17-16	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, ermA, tetEfflux, fosB	Rhine-Hesse	R20
t003	26-17-20-17-12-17-17-16	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, ermA, tetEfflux, fosB	Rhine-Hesse	R21
t504	26-17-20-17-12	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, ermA, tetEfflux, fosB	Rhine-Hesse	R18
t504	26-17-20-17-12	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, ermA, tetEfflux, fosB	Rhine-Hesse	R17
t003	26-17-20-17-12-17-17-16	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, ermA, aadD, tetEfflux, fosB	Rhine-Hesse	R19
t504	26-17-20-17-12	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, ermA, tetEfflux, fosB	Rhine-Hesse	R8
t887	26-17-20-17-12-17-17-12-17-17-16	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, ermA, tetEfflux, fosB	Rhine-Hesse	R7
t002	26-23-17-34-17-20-17-12-17-16	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, ermA, tetEfflux, fosB	Rhine-Hesse	R11
t003	26-17-20-17-12-17-17-16	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, ermA, tetEfflux, fosB	Rhine-Hesse	R12
t003	26-17-20-17-12-17-17-16	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, ermA, tetEfflux, fosB	Rhine-Hesse	R10
t003	26-17-20-17-12-17-17-16	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, ermA, tetEfflux, fosB	Rhine-Hesse	R9
t003	26-17-20-17-12-17-17-16	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, aadD, ermA, tetEfflux, fosB	Rhine-Hesse	R3
t045	26-17-20-17-12-17-16	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, aadD, ermA, tetEfflux, fosB	Rhine-Hesse	R14
t003	26-17-20-17-12-17-17-16	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, aadD, ermA, tetEfflux, fosB	Rhine-Hesse	R13
t010	26-17-34-17-20-17-12-17-16	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, aadD, ermA, tetEfflux, fosB	Rhine-Hesse	R5
t003	26-17-20-17-12-17-17-16	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, aadD, ermA, tetEfflux, fosB	Rhine-Hesse	R6
t504	26-17-20-17-12	CC5	II	II	sea, sed, sej, ser, egc	blaZ/I/R, aadD, ermA, tetEfflux, fosB	Rhine-Hesse	R4
t003	26-17-20-17-12-17-17-16	CC5	II	II	sed, sej, ser, egc	blaZ/I/R, ermA, tetEfflux, fosB	Rhine-Hesse	R39
t003	26-17-20-17-12-17-17-16	CC5	II	II	sed, sej, ser, egc	blaZ/I/R, ermA, tetEfflux, fosB	Rhine-Hesse	R38
t003	26-17-20-17-12-17-17-16	CC5	II	II	sed, sej, ser, egc	blaZ/I/R, ermA, tetEfflux, fosB	Rhine-Hesse	R1
t003	26-17-20-17-12-17-17-16	CC5	II	II	sed, sej, ser, egc	blaZ/I/R, ermA, aadD, fosB	Rhine-Hesse	R35
t003	26-17-20-17-12-17-17-16	CC5	II	II	sed, sej, ser, egc	blaZ/I/R, ermA, aadD, fosB	Rhine-Hesse	R36
t003	26-17-20-17-12-17-17-16	CC5	II	II	sed, sej, ser, egc	blaZ/I/R, ermA, aadD, tetEfflux, fosB	Rhine-Hesse	R37
t003	26-17-20-17-12-17-17-16	CC5	II	II	sed, sej, ser, egc	blaZ/I/R, ermA, tetEfflux, fosB	Rhine-Hesse	R2
t003	26-17-20-17-12-17-17-16	CC5	II	II	sed, sej, ser, egc	blaZ/I/R, ermA, aadD, tetEfflux, fosB	Rhine-Hesse	R34
t003	26-17-20-17-12-17-17-16	CC5	II	II	sea, egc	ermA, aacA-aphD, aadD, dfrA, tetEfflux, fosB, qacC	Rhine-Hesse	R25
t1079	17-20-17-12-17-17-16	CC5	II	II	sea, sej, ser, egc	ermA, aadD, tetEfflux, fosB	Rhine-Hesse	R26
t003	26-17-20-17-12-17-17-16	CC5	II	II	sea, egc	ermA, aadD, tetEfflux, fosB	Rhine-Hesse	R27
t481	26-17-20-17-12-17-17-16	CC5	II	II	egc	ermA, aadD, tetEfflux, fosB	Rhine-Hesse	R29
t003	26-17-20-17-12-17-17-16	CC5	II	II	egc	ermA, aadD, tetEfflux, fosB	Rhine-Hesse	R30
t3195	26-17-20-17-17-12-17-17-16	CC5	II	II	egc	ermA, aadD, tetEfflux, fosB	Rhine-Hesse	R28
t003	26-17-20-17-12-17-17-16	CC5	II	II	egc	ermA, tetEfflux, fosB	Rhine-Hesse	R31
t003	26-17-20-17-12-17-17-16	CC5	II	II	egc	ermA, tetEfflux, fosB	Rhine-Hesse	R33
t003	26-17-20-17-12-17-17-16	CC5	II	II	egc	ermA, aadD, tetEfflux, fosB	Rhine-Hesse	R32
t078	04-21-12-41-20-17-12-12-17	none	-	I	egc	blaZ/I/R, tetEfflux, fosB	Rhine-Hesse	S27
t003	26-17-20-17-12-17-17-16	CC5	II	II	sed, sej, ser, egc	blaZ/I/R, ermA, fosB	-	R40
t010	26-17-34-17-20-17-12-17-16	CC5	IV	II	egc	blaZ/I/R, aadD, fosB	not identified	R41
t493	04-34-17-66-32-17-23-24	CC5	-	II	sea, sed, sej, ser, egc	blaZ/I/R, ermA, tetEfflux, fosB	not identified	S43
t002	26-23-17-34-17-20-17-12-17-16	CC5	-	II	egc	tetEfflux, fosB	-	S42
t011	08-16-02-25-34-24-25	CC398	V	I	-	blaZ/I/R, tetK, tetM, tetEfflux	-	R44
t571	08-16-02-25-02-25-34-25	CC398	-	I	egc	blaZ/I/R, tetEfflux	Dutch-pig	S46
t005	26-23-13-23-31-05-17-25-17-25-16-28	CC22	-	I	egc	blaZ/I/R	-	S26
t022	26-23-13-23-31-29-17-31-29-17-25-17-25-16-28	CC22	IV	I	sec, sel, egc	blaZ/I/R	-	R46
t625	26-23-17-25-17-25-16-28	CC22	-	I	egc	-	Barnim	S25
t310	26-23-31-05-17-25-17-25-16-28	CC22	-	I	egc	-	-	S24
t018	15-12-16-02-16-02-25-17-24-24-24	CC15	-	I	-	blaZ/I/R, tetEfflux, fosB	-	S38
t084	07-23-12-34-34-12-12-23-02-12-23	CC15	-	I	-	blaZ/I/R, tetK, tetEfflux, fosB	-	S34
t8786	07-31-12-34-34-12-12-23-02-12-23	CC15	-	I	-	tetEfflux, fosB	-	S35
t084	07-23-12-34-34-12-12-23-02-12-23	CC15	-	I	-	tetEfflux, fosB	-	S37
t306	26-23-17-34-17-20-17-12-17-17-16	CC15	-	I	-	tetEfflux, fosB	-	S36
t267	07-23-12-21-17-34-34-34-33-34	CC97	-	I	-	blaZ/I/R, tetEfflux	-	S33
t267	07-23-12-21-17-34-34-34-33-34	CC97	-	I	-	blaZ/I/R, tetEfflux	-	S31
t267	07-23-12-21-17-34-34-34-33-34	CC97	-	I	-	blaZ/I/R, msrA, tetEfflux	-	S32
t131	07-23-12-34-33-34	CC97	-	I	-	tetEfflux	-	S29
t8830	07-23-12-21-17-17-34-34-33-34	CC97	-	I	-	blaZ/I/R, tetEfflux	-	S30
t008	11-19-12-21-17-34-24-34-22-25	CC8	-	I	sec, sel, egc	tetEfflux	-	S28
t008	11-19-12-21-17-34-24-34-22-25	CC8	-	I	sea, sed, sej, ser	blaZ/I/R, aadD, tetEfflux, fosB	-	R42
t008	11-19-12-21-17-34-24-34-22-25	CC8	-	I	sea	blaZ/I/R, ermA, aadD, tetEfflux, fosB, qacA	Lyon	R43
t8863	15-17-13-13-13-34-33-34	CC78	-	III	-	blaZ/I/R, tetEfflux	Lyon	S39
t8864	336-12-16-02-24	CC1	-	III	sea, sah	blaZ/I/R, tetEfflux	-	S40
t4044	04-20-17-12-17-17-17	CC101	-	I	-	tetEfflux, fosB	-	S45
t091	07-23-21-17-34-12-23-02-12-23	CC7	-	I	sea	tetEfflux	-	S44
t091	07-23-21-17-34-12-23-02-12-23	CC7	-	I	sea	blaZ/I/R, tetEfflux	-	S41
t584	15-16-02-16-02-25-17-24-24	CC30	-	II	tst-1, egc	blaZ/I/R, tetEfflux, fosB	-	S1
t273	07-23-21-17-13-34-16-34-33-13	CC30	-	II	tst-1, sea, egc	blaZ/I/R, linA, aphD, aadD, mupR, tetK, tetEfflux, fosB	-	S8
t012	15-12-16-02-16-02-25-17-24-24	CC30	-	II	egc	tetEfflux, fosB	-	S10
t8831	15-12-16-02-170-02-25-323-24	CC30	-	II	tst-1, egc	blaZ/I/R, tetEfflux, fosB	-	S9
t012	15-12-16-02-16-02-25-17-24-24	CC30	-	II	tst-1, egc	blaZ/I/R, tetEfflux, fosB	-	S3
t012	15-12-16-02-16-02-25-17-24-24	CC30	-	II	tst-1, egc	blaZ/I/R, tetEfflux, fosB	-	S2
t019	08-16-02-16-02-25-17-24	CC30	-	II	tst-1, sea, egc	blaZ/I/R, tetEfflux, fosB	-	S4
t012	15-12-16-02-16-02-25-17-24-24	CC30	-	II	sea, egc	blaZ/I/R, tetEfflux, fosB	-	S7
t012	15-12-16-02-16-02-25-17-24-24	CC30	-	II	tst-1, sea, egc	blaZ/I/R, tetEfflux, fosB	-	S6
t012	15-12-16-02-16-02-25-17-24-24	CC30	-	II	tst-1, sea, egc	blaZ/I/R, ermA, tetEfflux, fosB	-	S5
t1689	07-23-21-17-34-66-23-02-12-23	CC45	-	I	egc	tetEfflux	-	S12
t040	09-02-16-13-17-34-16-34	CC45	-	I	egc	ermA, tetEfflux	-	S11
t015	08-16-02-16-34-13-17-34-16-34	CC45	-	I	sec, sel, egc	blaZ/I/R, Q6GD50, tetEfflux	-	R45
t015	08-16-02-16-34-13-17-34-16-34	CC45	-	I	sec, sel, egc	blaZ/I/R, tetEfflux	-	S13
t073	08-16-02-16-13-17-34-16-34	CC45	-	I	egc	blaZ/I/R, tetEfflux	-	S14
t050	08-16-02-16-34-34-17-34-16-34	CC45	-	I	sec, sel, egc	blaZ/I/R, tetEfflux	-	S18
t015	08-16-02-16-34-13-17-34-16-34	CC45	-	I	sec, sel, egc	blaZ/I/R, tetEfflux	-	S17
t620	08-16-02-16-34-13-17-34	CC45	-	I	sec, sel, egc	blaZ/I/R, tetEfflux	-	S15
t015	08-16-02-16-34-13-17-34-16-34	CC45	-	I	sec, sel, egc	blaZ/I/R, tetEfflux	-	S16
t026	08-16-34	CC45	-	I	egc	tetEfflux	-	S23
t015	08-16-02-16-34-13-17-34-16-34	CC45	-	I	sec, sel, egc	tetEfflux	-	S22
t339	08-16-02-16	CC45	-	I	tst-1, sec, sel, egc	tetEfflux	-	S21
t2239	08-16-02-16-13-13-34-16-34	CC45	-	I	sec, sel, egc	tetEfflux	-	S20
t015	08-16-02-16-34-13-17-34-16-34	CC45	-	nd	sec, sel, egc	-	-	S19

FIGURE A.1: Clustering

Appendix B

Co-methylation supplementary material (chapter 4)

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
1	CD1C	hsa04640:Hematopoietic cell lineage,
1	CD48	hsa04650:Natural killer cell mediated cytotoxicity,
1	GALNT8	hsa00512:O-Glycan biosynthesis,
1	EPX	hsa05310:Asthma,
1	FSHR	hsa04080:Neuroactive ligand-receptor interaction,
1	GRM8	hsa04080:Neuroactive ligand-receptor interaction,
1	IFNA8	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04140:Regulation of autophagy,
		hsa04612:Antigen processing and presentation,
		hsa04620:Toll-like receptor signaling pathway,
		hsa04622:RIG-I-like receptor signaling pathway,
		hsa04623:Cytosolic DNA-sensing pathway,
		hsa04630:Jak-STAT signaling pathway,
		hsa04650:Natural killer cell mediated cytotoxicity,
		hsa05320:Autoimmune thyroid disease,
1	MASP1	hsa04610:Complement and coagulation cascades,
1	MBL2	hsa04610:Complement and coagulation cascades,
1	TAAR5	hsa04080:Neuroactive ligand-receptor interaction,
1	VNN1	hsa00770:Pantothenate and CoA biosynthesis,
2	AIFM1	hsa04210:Apoptosis,
2	COX7B	hsa00190:Oxidative phosphorylation,
		hsa04260:Cardiac muscle contraction,
		hsa05010:Alzheimer's disease,
		hsa05012:Parkinson's disease,
		hsa05016:Huntington's disease,
2	PDHA1	hsa00010:Glycolysis \Gluconeogenesis,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
		hsa00020:Citrate cycle (TCA cycle),
		hsa00290:Valine, leucine and isoleucine biosynthesis,
		hsa00620:Pyruvate metabolism,
		hsa00650:Butanoate metabolism,
3	CYSLTR2	hsa04020:Calcium signaling pathway,
		hsa04080:Neuroactive ligand-receptor interaction,
3	GNRH1	hsa04912:GnRH signaling pathway,
3	LEFTY1	hsa04350:TGF-beta signaling pathway,
3	PLXNB1	hsa04360:Axon guidance,
3	PPP1R3A	hsa04910:Insulin signaling pathway,
4	ARHGEF4	hsa04810:Regulation of actin cytoskeleton,
4	CCL7	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04062:Chemokine signaling pathway,
		hsa04621:NOD-like receptor signaling pathway,
4	GABRA5	hsa04080:Neuroactive ligand-receptor interaction,
4	OR5P2	hsa04740:Olfactory transduction,
5	HTR4	hsa04020:Calcium signaling pathway,
		hsa04080:Neuroactive ligand-receptor interaction,
5	BMP8A	hsa04340:Hedgehog signaling pathway,
		hsa04350:TGF-beta signaling pathway,
5	EREG	hsa04012:ErbB signaling pathway,
5	EPB41L3	hsa04530:Tight junction,
5	JAM3	hsa04514:Cell adhesion molecules (CAMs),
		hsa04530:Tight junction,
		hsa04670:Leukocyte transendothelial migration,
		hsa05120:Epithelial cell signaling in Helicobacter pylori infection,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
5	LPL	hsa00561:Glycerolipid metabolism,
		hsa03320:PPAR signaling pathway,
		hsa05010:Alzheimer's disease,
5	PDX1	hsa04930:Type II diabetes mellitus,
		hsa04950:Maturity onset diabetes of the young,
5	PLA2G7	hsa00565:Ether lipid metabolism,
6	CD244	hsa04650:Natural killer cell mediated cytotoxicity,
6	ALDOB	hsa00010:Glycolysis / Gluconeogenesis,
		hsa00030:Pentose phosphate pathway,
		hsa00051:Fructose and mannose metabolism,
6	AIRE	hsa04120:Ubiquitin mediated proteolysis,
		hsa05340:Primary immunodeficiency,
6	CACNG5	hsa04010:MAPK signaling pathway,
		hsa04260:Cardiac muscle contraction,
		hsa05410:Hypertrophic cardiomyopathy (HCM),
		hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC),
		hsa05414:Dilated cardiomyopathy,
6	CES7	hsa00983:Drug metabolism,
6	CCL8	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04062:Chemokine signaling pathway,
		hsa04621:NOD-like receptor signaling pathway,
6	C1QB	hsa04610:Complement and coagulation cascades,
		hsa05020:Prion diseases,
		hsa05322:Systemic lupus erythematosus,
6	C9	hsa04610:Complement and coagulation cascades,
		hsa05020:Prion diseases,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
		hsa05322:Systemic lupus erythematosus,
6	ENTPD1	hsa00230:Purine metabolism,
		hsa00240:Pyrimidine metabolism,
6	KLK3	hsa05200:Pathways in cancer,
		hsa05215:Prostate cancer,
6	MYH6	hsa04260:Cardiac muscle contraction,
		hsa04530:Tight junction,
		hsa05410:Hypertrophic cardiomyopathy (HCM),
		hsa05414:Dilated cardiomyopathy,
		hsa05416:Viral myocarditis,
6	OR10H3	hsa04740:Olfactory transduction,
6	PRODH2	hsa00330:Arginine and proline metabolism,
6	PRSS1	hsa04080:Neuroactive ligand-receptor interaction,
6	RIPK3	hsa04623:Cytosolic DNA-sensing pathway,
7	LIMK1	hsa04360:Axon guidance,
		hsa04666:Fc gamma R-mediated phagocytosis,
		hsa04810:Regulation of actin cytoskeleton,
7	B4GALT6	hsa00600:Sphingolipid metabolism,
7	ALDOC	hsa00010:Glycolysis / Gluconeogenesis,
		hsa00030:Pentose phosphate pathway,
		hsa00051:Fructose and mannose metabolism,
7	EGFR	hsa04010:MAPK signaling pathway,
		hsa04012:ErbB signaling pathway,
		hsa04020:Calcium signaling pathway,
		hsa04060:Cytokine-cytokine receptor interaction,
		hsa04144:Endocytosis,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
		hsa04320:Dorso-ventral axis formation,
		hsa04510:Focal adhesion,
		hsa04520:Adherens junction,
		hsa04540:Gap junction,
		hsa04810:Regulation of actin cytoskeleton,
		hsa04912:GnRH signaling pathway,
		hsa05120:Epithelial cell signaling in Helicobacter pylori infection,
		hsa05200:Pathways in cancer,
		hsa05210:Colorectal cancer,
		hsa05212:Pancreatic cancer,
		hsa05213:Endometrial cancer,
		hsa05214:Glioma,
		hsa05215:Prostate cancer,
		hsa05218:Melanoma,
		hsa05219:Bladder cancer,
		hsa05223:Non-small cell lung cancer,
7	FADS2	hsa00592:alpha-Linolenic acid metabolism,
		hsa01040:Biosynthesis of unsaturated fatty acids,
		hsa03320:PPAR signaling pathway,
7	HIST1H3J	hsa05322:Systemic lupus erythematosus,
7	MAGI2	hsa04530:Tight junction,
7	RPL31	hsa03010:Ribosome,
7	SCNN1B	hsa04742:Taste transduction,
		hsa04960:Aldosterone-regulated sodium reabsorption,
7	TUBB6	hsa04540:Gap junction,
		hsa05130:Pathogenic Escherichia coli infection,

Continued

TABLE B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
8	HTR1B	hsa04080:Neuroactive ligand-receptor interaction,
8	HNF1B	hsa04950:Maturity onset diabetes of the young,
8	ACTN2	hsa04510:Focal adhesion,
		hsa04520:Adherens junction,
		hsa04530:Tight junction,
		hsa04670:Leukocyte transendothelial migration,
		hsa04810:Regulation of actin cytoskeleton,
		hsa05322:Systemic lupus erythematosus,
		hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC),
8	APC2	hsa04310:Wnt signaling pathway,
		hsa04810:Regulation of actin cytoskeleton,
		hsa05200:Pathways in cancer,
		hsa05210:Colorectal cancer,
		hsa05213:Endometrial cancer,
		hsa05217:Basal cell carcinoma,
8	ALDH1A2	hsa00830:Retinol metabolism,
8	BDNF	hsa04010:MAPK signaling pathway,
		hsa04722:Neurotrophin signaling pathway,
		hsa05016:Huntington's disease,
8	CACNA1A	hsa04010:MAPK signaling pathway,
		hsa04020:Calcium signaling pathway,
		hsa04730:Long-term depression,
		hsa04742:Taste transduction,
		hsa04930:Type II diabetes mellitus,
8	CCNA1	hsa04110:Cell cycle,
		hsa04914:Progesterone-mediated oocyte maturation,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
		hsa05200:Pathways in cancer,
		hsa05221:Acute myeloid leukemia,
8	DRD5	hsa04020:Calcium signaling pathway,
		hsa04080:Neuroactive ligand-receptor interaction,
8	EPO	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04630:Jak-STAT signaling pathway,
		hsa04640:Hematopoietic cell lineage,
8	FOXA2	hsa04950:Maturity onset diabetes of the young,
8	GHSR	hsa04080:Neuroactive ligand-receptor interaction,
8	NEUROD1	hsa04950:Maturity onset diabetes of the young,
8	NOS1	hsa00330:Arginine and proline metabolism,
		hsa04020:Calcium signaling pathway,
		hsa04730:Long-term depression,
		hsa05010:Alzheimer's disease,
		hsa05014:Amyotrophic lateral sclerosis (ALS),
8	OPRM1	hsa04080:Neuroactive ligand-receptor interaction,
8	PTPRN	hsa04940:Type I diabetes mellitus,
8	TNFRSF8	hsa04060:Cytokine-cytokine receptor interaction,
8	WNT2	hsa04310:Wnt signaling pathway,
		hsa04340:Hedgehog signaling pathway,
		hsa04916:Melanogenesis,
		hsa05200:Pathways in cancer,
		hsa05217:Basal cell carcinoma,
9	CD1A	hsa04640:Hematopoietic cell lineage,
9	CD1D	hsa04640:Hematopoietic cell lineage,
9	CD1E	hsa04640:Hematopoietic cell lineage,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
9	CDH4	hsa04514:Cell adhesion molecules (CAMs),
9	CCL11	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04062:Chemokine signaling pathway,
		hsa04621:NOD-like receptor signaling pathway,
		hsa05310:Asthma,
9	LAMA3	hsa04510:Focal adhesion,
		hsa04512:ECM-receptor interaction,
		hsa05200:Pathways in cancer,
		hsa05222:Small cell lung cancer,
9	MYH4	hsa04530:Tight junction,
		hsa05416:Viral myocarditis,
9	OR1G1	hsa04740:Olfactory transduction,
9	OR12D3	hsa04740:Olfactory transduction,
9	PAX4	hsa04950:Maturity onset diabetes of the young,
10	CD34	hsa04514:Cell adhesion molecules (CAMs),
		hsa04640:Hematopoietic cell lineage,
10	ST6GALNAC1	hsa00512:O-Glycan biosynthesis,
10	CA9	hsa00910:Nitrogen metabolism,
10	CCKAR	hsa04020:Calcium signaling pathway,
		hsa04080:Neuroactive ligand-receptor interaction,
10	F2RL3	hsa04080:Neuroactive ligand-receptor interaction,
10	C1S	hsa04610:Complement and coagulation cascades,
		hsa05322:Systemic lupus erythematosus,
10	DOCK2	hsa04062:Chemokine signaling pathway,
		hsa04666:Fc gamma R-mediated phagocytosis,
10	GABRD	hsa04080:Neuroactive ligand-receptor interaction,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
10	GNA13	hsa04270:Vascular smooth muscle contraction,
		hsa04730:Long-term depression,
		hsa04810:Regulation of actin cytoskeleton,
10	IGF1	hsa04114:Oocyte meiosis,
		hsa04115:p53 signaling pathway,
		hsa04150:mTOR signaling pathway,
		hsa04510:Focal adhesion,
		hsa04730:Long-term depression,
		hsa04914:Progesterone-mediated oocyte maturation,
		hsa04960:Aldosterone-regulated sodium reabsorption,
		hsa05200:Pathways in cancer,
		hsa05214:Glioma,
		hsa05215:Prostate cancer,
		hsa05218:Melanoma,
		hsa05410:Hypertrophic cardiomyopathy (HCM),
		hsa05414:Dilated cardiomyopathy,
10	LY96	hsa04620:Toll-like receptor signaling pathway,
		hsa05130:Pathogenic Escherichia coli infection,
10	NPFFR2	hsa04080:Neuroactive ligand-receptor interaction,
10	PITX2	hsa04350:TGF-beta signaling pathway,
10	PCYT1B	hsa00564:Glycerophospholipid metabolism,
10	PDE4C	hsa00230:Purine metabolism,
10	P4HA3	hsa00330:Arginine and proline metabolism,
10	POMC	hsa04916:Melanogenesis,
		hsa04920:Adipocytokine signaling pathway,
10	SGCD	hsa05410:Hypertrophic cardiomyopathy (HCM),

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
		hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC),
		hsa05414:Dilated cardiomyopathy,
		hsa05416:Viral myocarditis,
10	TNFRSF17	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04672:Intestinal immune network for IgA production,
11	HAAO	hsa00380:Tryptophan metabolism,
11	CD38	hsa00760:Nicotinate and nicotinamide metabolism,
		hsa04020:Calcium signaling pathway,
		hsa04640:Hematopoietic cell lineage,
11	AKR1B1	hsa00040:Pentose and glucuronate interconversions,
		hsa00051:Fructose and mannose metabolism,
		hsa00052:Galactose metabolism,
		hsa00561:Glycerolipid metabolism,
		hsa00620:Pyruvate metabolism,
11	CNTN2	hsa04514:Cell adhesion molecules (CAMs),
11	FABP5	hsa03320:PPAR signaling pathway,
11	GRIA1	hsa04080:Neuroactive ligand-receptor interaction,
		hsa04720:Long-term potentiation,
		hsa04730:Long-term depression,
		hsa05014:Amyotrophic lateral sclerosis (ALS),
11	GNA15	hsa04020:Calcium signaling pathway,
11	GUCY2D	hsa00230:Purine metabolism,
		hsa04740:Olfactory transduction,
11	IL23A	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04630:Jak-STAT signaling pathway,
11	MAT1A	hsa00270:Cysteine and methionine metabolism,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
		hsa00450:Selenoamino acid metabolism,
11	NPY	hsa04920:Adipocytokine signaling pathway,
11	NODAL	hsa04350:TGF-beta signaling pathway,
11	SIPA1	hsa04670:Leukocyte transendothelial migration,
11	SLC8A2	hsa04020:Calcium signaling pathway,
11	TNFRSF10D	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04210:Apoptosis,
		hsa04650:Natural killer cell mediated cytotoxicity,
11	TNFRSF1B	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04920:Adipocytokine signaling pathway,
		hsa05014:Amyotrophic lateral sclerosis (ALS),
12	CLDN8	hsa04514:Cell adhesion molecules (CAMs),
		hsa04530:Tight junction,
		hsa04670:Leukocyte transendothelial migration,
12	KLHL13	hsa04120:Ubiquitin mediated proteolysis,
12	TNFRSF9	hsa04060:Cytokine-cytokine receptor interaction,
12	VDAC1	hsa04020:Calcium signaling pathway,
		hsa05012:Parkinson's disease,
		hsa05016:Huntington's disease,
13	CFB	hsa04610:Complement and coagulation cascades,
13	CYFIP2	hsa04810:Regulation of actin cytoskeleton,
13	IL18	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04621:NOD-like receptor signaling pathway,
		hsa04623:Cytosolic DNA-sensing pathway,
13	HLA-DRA	hsa04514:Cell adhesion molecules (CAMs),
		hsa04612:Antigen processing and presentation,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
		hsa04640:Hematopoietic cell lineage,
		hsa04672:Intestinal immune network for IgA production,
		hsa04940:Type I diabetes mellitus,
		hsa05310:Asthma,
		hsa05320:Autoimmune thyroid disease,
		hsa05322:Systemic lupus erythematosus,
		hsa05330:Allograft rejection,
		hsa05332:Graft-versus-host disease,
		hsa05416:Viral myocarditis,
13	RHOH	hsa04670:Leukocyte transendothelial migration,
13	TJP3	hsa04530:Tight junction,
14	CD40	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04514:Cell adhesion molecules (CAMs),
		hsa04620:Toll-like receptor signaling pathway,
		hsa04672:Intestinal immune network for IgA production,
		hsa05310:Asthma,
		hsa05320:Autoimmune thyroid disease,
		hsa05322:Systemic lupus erythematosus,
		hsa05330:Allograft rejection,
		hsa05340:Primary immunodeficiency,
		hsa05416:Viral myocarditis,
14	CTPS	hsa00240:Pyrimidine metabolism,
14	CHST3	hsa00532:Chondroitin sulfate biosynthesis,
14	CXCL12	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04062:Chemokine signaling pathway,
		hsa04360:Axon guidance,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
		hsa04670:Leukocyte transendothelial migration,
		hsa04672:Intestinal immune network for IgA production,
14	CDO1	hsa00270:Cysteine and methionine metabolism,
		hsa00430:Taurine and hypotaurine metabolism,
14	HIST1H4L	hsa05322:Systemic lupus erythematosus,
14	IRAK3	hsa04210:Apoptosis,
		hsa04722:Neurotrophin signaling pathway,
14	NPR2	hsa00230:Purine metabolism,
		hsa04270:Vascular smooth muscle contraction,
14	PLAT	hsa04610:Complement and coagulation cascades,
14	VAV1	hsa04062:Chemokine signaling pathway,
		hsa04510:Focal adhesion,
		hsa04650:Natural killer cell mediated cytotoxicity,
		hsa04660:T cell receptor signaling pathway,
		hsa04662:B cell receptor signaling pathway,
		hsa04664:Fc epsilon RI signaling pathway,
		hsa04666:Fc gamma R-mediated phagocytosis,
		hsa04670:Leukocyte transendothelial migration,
		hsa04810:Regulation of actin cytoskeleton,
15	C7orf16	hsa04730:Long-term depression,
15	C6	hsa04610:Complement and coagulation cascades,
		hsa05020:Prion diseases,
		hsa05322:Systemic lupus erythematosus,
15	GRM5	hsa04020:Calcium signaling pathway,
		hsa04080:Neuroactive ligand-receptor interaction,
		hsa04540:Gap junction,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
		hsa04720:Long-term potentiation,
		hsa04730:Long-term depression,
		hsa05016:Huntington's disease,
15	GYS2	hsa00500:Starch and sucrose metabolism,
		hsa04910:Insulin signaling pathway,
15	IFNG	hsa03050:Proteasome,
		hsa04060:Cytokine-cytokine receptor interaction,
		hsa04140:Regulation of autophagy,
		hsa04350:TGF-beta signaling pathway,
		hsa04630:Jak-STAT signaling pathway,
		hsa04650:Natural killer cell mediated cytotoxicity,
		hsa04660:T cell receptor signaling pathway,
		hsa04940:Type I diabetes mellitus,
		hsa05322:Systemic lupus erythematosus,
		hsa05330:Allograft rejection,
		hsa05332:Graft-versus-host disease,
15	OR5I1	hsa04740:Olfactory transduction,
15	PRKG2	hsa04540:Gap junction,
		hsa04730:Long-term depression,
		hsa04740:Olfactory transduction,
15	SGCG	hsa05410:Hypertrophic cardiomyopathy (HCM),
		hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC),
		hsa05414:Dilated cardiomyopathy,
		hsa05416:Viral myocarditis,
16	ACOX2	hsa00120:Primary bile acid biosynthesis,
		hsa03320:PPAR signaling pathway,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
16	CARD11	hsa04660:T cell receptor signaling pathway,
		hsa04662:B cell receptor signaling pathway,
16	GRIK2	hsa04080:Neuroactive ligand-receptor interaction,
16	GDA	hsa00230:Purine metabolism,
16	LCP2	hsa04650:Natural killer cell mediated cytotoxicity,
		hsa04660:T cell receptor signaling pathway,
		hsa04664:Fc epsilon RI signaling pathway,
16	MCHR2	hsa04080:Neuroactive ligand-receptor interaction,
16	MAOA	hsa00260:Glycine, serine and threonine metabolism,
		hsa00330:Arginine and proline metabolism,
		hsa00340:Histidine metabolism,
		hsa00350:Tyrosine metabolism,
		hsa00360:Phenylalanine metabolism,
		hsa00380:Tryptophan metabolism,
		hsa00982:Drug metabolism,
16	PRLHR	hsa04080:Neuroactive ligand-receptor interaction,
16	RUNX1T1	hsa05200:Pathways in cancer,
		hsa05221:Acute myeloid leukemia,
16	RYR2	hsa04020:Calcium signaling pathway,
		hsa04260:Cardiac muscle contraction,
		hsa05410:Hypertrophic cardiomyopathy (HCM),
		hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC),
		hsa05414:Dilated cardiomyopathy,
17	H2AFY	hsa05322:Systemic lupus erythematosus,
17	RAB4A	hsa04144:Endocytosis,
17	ADCY4	hsa00230:Purine metabolism,

Continued

TABLE B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
		hsa04020:Calcium signaling pathway,
		hsa04062:Chemokine signaling pathway,
		hsa04114:Oocyte meiosis,
		hsa04270:Vascular smooth muscle contraction,
		hsa04540:Gap junction,
		hsa04742:Taste transduction,
		hsa04912:GnRH signaling pathway,
		hsa04914:Progesterone-mediated oocyte maturation,
		hsa04916:Melanogenesis,
		hsa05414:Dilated cardiomyopathy,
17	CDH5	hsa04514:Cell adhesion molecules (CAMs),
		hsa04670:Leukocyte transendothelial migration,
17	DNM3	hsa04144:Endocytosis,
		hsa04666:Fc gamma R-mediated phagocytosis,
17	FMO2	hsa00982:Drug metabolism,
17	LAMB3	hsa04510:Focal adhesion,
		hsa04512:ECM-receptor interaction,
		hsa05200:Pathways in cancer,
		hsa05222:Small cell lung cancer,
17	MMP7	hsa04310:Wnt signaling pathway,
17	MGST1	hsa00480:Glutathione metabolism,
		hsa00980:Metabolism of xenobiotics by cytochrome P450,
		hsa00982:Drug metabolism,
17	RDH5	hsa00830:Retinol metabolism,
19	ABO	hsa00601:Glycosphingolipid biosynthesis,
19	ST8SIA5	hsa00604:Glycosphingolipid biosynthesis,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
19	ALPL	hsa00790:Folate biosynthesis,
19	ALOX15	hsa00590:Arachidonic acid metabolism,
		hsa00591:Linoleic acid metabolism,
19	CRHR2	hsa04080:Neuroactive ligand-receptor interaction,
19	ITGA4	hsa04510:Focal adhesion,
		hsa04512:ECM-receptor interaction,
		hsa04514:Cell adhesion molecules (CAMs),
		hsa04640:Hematopoietic cell lineage,
		hsa04670:Leukocyte transendothelial migration,
		hsa04672:Intestinal immune network for IgA production,
		hsa04810:Regulation of actin cytoskeleton,
		hsa05410:Hypertrophic cardiomyopathy (HCM),
		hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC),
		hsa05414:Dilated cardiomyopathy,
20	ATP6V0C	hsa00190:Oxidative phosphorylation,
		hsa04142:Lysosome,
		hsa05110:Vibrio cholerae infection,
		hsa05120:Epithelial cell signaling in Helicobacter pylori infection,
20	CD8B	hsa04514:Cell adhesion molecules (CAMs),
		hsa04612:Antigen processing and presentation,
		hsa04640:Hematopoietic cell lineage,
		hsa04660:T cell receptor signaling pathway,
		hsa05340:Primary immunodeficiency,
20	RAB11FIP4	hsa04144:Endocytosis,
20	ACSS1	hsa00010:Glycolysis / Gluconeogenesis,
		hsa00620:Pyruvate metabolism,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
		hsa00640:Propanoate metabolism,
20	APC	hsa04310:Wnt signaling pathway,
		hsa04810:Regulation of actin cytoskeleton,
		hsa05200:Pathways in cancer,
		hsa05210:Colorectal cancer,
		hsa05213:Endometrial cancer,
		hsa05217:Basal cell carcinoma,
20	ADA	hsa00230:Purine metabolism,
		hsa05340:Primary immunodeficiency,
20	ALOX15B	hsa00590:Arachidonic acid metabolism,
20	CXCL6	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04062:Chemokine signaling pathway,
20	CX3CL1	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04062:Chemokine signaling pathway,
20	C5AR1	hsa04080:Neuroactive ligand-receptor interaction,
		hsa04610:Complement and coagulation cascades,
20	HIST1H4F	hsa05322:Systemic lupus erythematosus,
20	MMP2	hsa04670:Leukocyte transendothelial migration,
		hsa04912:GnRH signaling pathway,
		hsa05200:Pathways in cancer,
		hsa05219:Bladder cancer,
21	CYBA	hsa04670:Leukocyte transendothelial migration,
21	EXTL2	hsa00534:Heparan sulfate biosynthesis,
21	GSTM2	hsa00480:Glutathione metabolism,
		hsa00980:Metabolism of xenobiotics by cytochrome P450,
		hsa00982:Drug metabolism,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
21	HIST1H3I	hsa05322:Systemic lupus erythematosus,
21	MMP14	hsa04912:GnRH signaling pathway,
21	NCAM2	hsa04514:Cell adhesion molecules (CAMs),
		hsa05020:Prion diseases,
21	UCP1	hsa03320:PPAR signaling pathway,
		hsa05016:Huntington's disease,
22	CLCA1	hsa04740:Olfactory transduction,
22	F7	hsa04610:Complement and coagulation cascades,
22	EGF	hsa04010:MAPK signaling pathway,
		hsa04012:ErbB signaling pathway,
		hsa04060:Cytokine-cytokine receptor interaction,
		hsa04144:Endocytosis,
		hsa04510:Focal adhesion,
		hsa04540:Gap junction,
		hsa04810:Regulation of actin cytoskeleton,
		hsa05200:Pathways in cancer,
		hsa05212:Pancreatic cancer,
		hsa05213:Endometrial cancer,
		hsa05214:Glioma,
		hsa05215:Prostate cancer,
		hsa05218:Melanoma,
		hsa05219:Bladder cancer,
		hsa05223:Non-small cell lung cancer,
22	EPHX1	hsa00980:Metabolism of xenobiotics by cytochrome P450,
22	IL20	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04630:Jak-STAT signaling pathway,

Continued

TABLE B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
22	SCNN1A	hsa04742:Taste transduction,
		hsa04960:Aldosterone-regulated sodium reabsorption,
22	THPO	hsa04640:Hematopoietic cell lineage,
22	TNFSF18	hsa04060:Cytokine-cytokine receptor interaction,
23	DUSP16	hsa04010:MAPK signaling pathway,
23	LRDD	hsa04115:p53 signaling pathway,
23	NCL	hsa05130:Pathogenic Escherichia coli infection,
23	PPIL2	hsa04120:Ubiquitin mediated proteolysis,
23	SC4MOL	hsa00100:Steroid biosynthesis,
23	SOCS3	hsa04120:Ubiquitin mediated proteolysis,
		hsa04630:Jak-STAT signaling pathway,
		hsa04910:Insulin signaling pathway,
		hsa04920:Adipocytokine signaling pathway,
		hsa04930:Type II diabetes mellitus,
24	CHST2	hsa00533:Keratan sulfate biosynthesis,
24	CNTNAP2	hsa04514:Cell adhesion molecules (CAMs),
24	GRIN3A	hsa04080:Neuroactive ligand-receptor interaction,
25	BCL10	hsa04660:T cell receptor signaling pathway,
		hsa04662:B cell receptor signaling pathway,
25	DBH	hsa00350:Tyrosine metabolism,
26	F2RL1	hsa04080:Neuroactive ligand-receptor interaction,
26	EXTL1	hsa00534:Heparan sulfate biosynthesis,
26	FABP3	hsa03320:PPAR signaling pathway,
26	HIST1H3G	hsa05322:Systemic lupus erythematosus,
26	IFNGR2	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04630:Jak-STAT signaling pathway,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
		hsa04650:Natural killer cell mediated cytotoxicity,
26	LHCGR	hsa04020:Calcium signaling pathway,
		hsa04080:Neuroactive ligand-receptor interaction,
26	PGAM2	hsa00010:Glycolysis / Gluconeogenesis,
26	PIPOX	hsa00260:Glycine, serine and threonine metabolism,
		hsa00310:Lysine degradation,
26	PDCD1LG2	hsa04514:Cell adhesion molecules (CAMs),
26	PPP1R14A	hsa04270:Vascular smooth muscle contraction,
26	PTPRR	hsa04010:MAPK signaling pathway,
26	SREBF1	hsa04910:Insulin signaling pathway,
26	SV2A	hsa04512:ECM-receptor interaction,
26	TRIP10	hsa04910:Insulin signaling pathway,
26	UCK1	hsa00240:Pyrimidine metabolism,
		hsa00983:Drug metabolism,
27	CHRM5	hsa04020:Calcium signaling pathway,
		hsa04080:Neuroactive ligand-receptor interaction,
		hsa04810:Regulation of actin cytoskeleton,
27	COL5A2	hsa04510:Focal adhesion,
		hsa04512:ECM-receptor interaction,
27	COL11A1	hsa04510:Focal adhesion,
		hsa04512:ECM-receptor interaction,
27	CA5B	hsa00910:Nitrogen metabolism,
27	NTSR1	hsa04020:Calcium signaling pathway,
		hsa04080:Neuroactive ligand-receptor interaction,
27	KCNMB2	hsa04270:Vascular smooth muscle contraction,
27	SPP1	hsa04510:Focal adhesion,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
		hsa04512:ECM-receptor interaction,
		hsa04620:Toll-like receptor signaling pathway,
27	TAT	hsa00130:Ubiquinone and other terpenoid-quinone biosynthesis,
		hsa00270:Cysteine and methionine metabolism,
		hsa00350:Tyrosine metabolism,
		hsa00360:Phenylalanine metabolism,
		hsa00400:Phenylalanine, tyrosine and tryptophan biosynthesis,
28	NLRP3	hsa04621:NOD-like receptor signaling pathway,
28	CYP11B1	hsa00140:Steroid hormone biosynthesis,
		hsa00150:Androgen and estrogen metabolism,
28	FGF6	hsa04010:MAPK signaling pathway,
		hsa04810:Regulation of actin cytoskeleton,
		hsa05200:Pathways in cancer,
		hsa05218:Melanoma,
28	GRIK5	hsa04080:Neuroactive ligand-receptor interaction,
28	NOS3	hsa00330:Arginine and proline metabolism,
		hsa04020:Calcium signaling pathway,
		hsa04370:VEGF signaling pathway,
28	OR7A5	hsa04740:Olfactory transduction,
28	OR8B8	hsa04740:Olfactory transduction,
28	TPO	hsa00350:Tyrosine metabolism,
		hsa04060:Cytokine-cytokine receptor interaction,
		hsa04630:Jak-STAT signaling pathway,
		hsa04640:Hematopoietic cell lineage,
		hsa05320:Autoimmune thyroid disease,
29	HUWE1	hsa04120:Ubiquitin mediated proteolysis,

Continued

Table B.1: Pathway mapping of 29 clusters using DAVID. These clusters of genes are identical to clusters described in Table S1.

Cluster ID	Gene symbol	KEGG pathways
29	AMH	hsa04060:Cytokine-cytokine receptor interaction,
		hsa04350:TGF-beta signaling pathway,
29	CTNNBL1	hsa03040:Spliceosome,
29	COL3A1	hsa04510:Focal adhesion,
		hsa04512:ECM-receptor interaction,
29	MAP3K14	hsa04010:MAPK signaling pathway,
		hsa04210:Apoptosis,
		hsa04660:T cell receptor signaling pathway,
		hsa04672:Intestinal immune network for IgA production,
		hsa05120:Epithelial cell signaling in Helicobacter pylori infection,
29	PLA2G2E	hsa00564:Glycerophospholipid metabolism,
		hsa00565:Ether lipid metabolism,
		hsa00590:Arachidonic acid metabolism,
		hsa00591:Linoleic acid metabolism,
		hsa00592:alpha-Linolenic acid metabolism,
		hsa04010:MAPK signaling pathway,
		hsa04270:Vascular smooth muscle contraction,
		hsa04370:VEGF signaling pathway,
		hsa04664:Fc epsilon RI signaling pathway,
		hsa04730:Long-term depression,
		hsa04912:GnRH signaling pathway,
29	KCNB1	hsa04742:Taste transduction.

TABLE B.2: The 187 strongest correlations for pairs of genes with respect to their -values, obtained after three stage filtering. For these pairs of genes, we also computed the Pearson correlation of their expression values (fourth column). In case at least one out of two genes was not found in the gene expression sample, NA was assigned to their co-expression value. Genomic distance contains d for those pairs of genes, which are located on different chromosomes.

First gene	Second gene	Co-methylation	Co-expression	rfunSimAll	Genomic distance
SPRR1B	SPRR1A	0.872	0.584	0.96	48136
FCN2	FCN1	0.870	0.260	0.98	37462
CD244	CD48	0.866	0.713	0.85	151800
SPRR1B	SPRR4	0.862	0.101	NA	61112
TAS2R13	PRB4	0.859	0.456	0	401809
F7	TFF1	0.856	0.462	0.81	d
SH3TC2	SPARCL1	0.853	0.070	NA	d
ABCE1	SC4MOL	0.849	0.122	0.25	20230109
REG1B	REG1P	0.846	NA	0.92	67887
SPRR3	SPRR4	0.843	0.078	NA	30747
SPRR1B	SPRR2D	0.842	0.783	NA	10959
TAS2R13	SCOC	0.842	-0.060	NA	d
C3orf32	TFF1	0.840	0.060	NA	d
REG1B	REG1A	0.840	0.331	NA	31695
PGLYRP3	LOR	0.836	0.050	0.11	51646
SPRR1A	SPRR4	0.834	0.152	NA	12976
C1orf64	TFF1	0.832	0.415	NA	d
KRTAP8-1	KRTAP20-1	0.830	NA	NA	196749
TXLNA	SC4MOL	0.829	0.102	0.10	d
TFF1	TNFSF18	0.827	0.051	0.70	d
REG1P	REG3A	0.827	NA	0.92	3382
KRTAP8-1	KRTAP21-1	0.826	NA	NA	56305
KRTAP21-1	KRTAP20-1	0.824	NA	NA	140444

Continued

TABLE B.2: The 187 strongest correlations for pairs of genes with respect to their -values, obtained after three stage filtering. For these pairs of genes, we also computed the Pearson correlation of their expression values (fourth column). In case at least one out of two genes was not found in the gene expression sample, NA was assigned to their co-expression value. Genomic distance contains d for those pairs of genes, which are located on different chromosomes.

First gene	**Second gene**	**Co-methylation**	**Co-expression**	**rfunSimAll**	**Genomic distance**
CDH5	NDRG2	0.821	0.121	0.43	d
PPIL2	SC4MOL	0.820	0.112	0.16	d
HBB	HBE1	0.820	-0.018	0.92	41218
TXLNA	PPIL2	0.820	0.250	0.13	d
CD8A	PODN	0.817	NA	0.49	d
CD48	SLAMF1	0.816	0.851	0.79	64890
C3orf32	SCOC	0.815	0.022	NA	d
MYH1	MYH4	0.813	0.341	NA	45771
REG1P	REG1A	0.810	NA	NA	36192
SPRR2D	SPRR1A	0.809	0.662	NA	59095
DNM3	INA	0.808	-0.021	0.53	d
SCOC	TFF1	0.808	0.014	NA	d
TXLNA	ABCE1	0.807	-0.199	0.38	d
KRTAP20-1	KRTAP13-4	0.806	NA	NA	185787
TAS2R13	TNFSF18	0.806	-0.048	0.48	d
PODN	INA	0.805	-0.218	0.08	d
PRKCB	MSC	0.804	NA	0.67	d
TAS2R13	TFF1	0.804	-0.033	0.28	d
C3orf32	C1orf64	0.804	0.095	NA	d
SPRR2D	SPRR4	0.803	0.118	NA	72071
CD8A	ST8SIA5	0.802	NA	0.24	d
SERPINB12	SPP2	0.801	0.013	0.68	d
REG1B	MORC1	0.800	0.032	0.18	d

Continued

TABLE B.2: The 187 strongest correlations for pairs of genes with respect to their -values, obtained after three stage filtering. For these pairs of genes, we also computed the Pearson correlation of their expression values (fourth column). In case at least one out of two genes was not found in the gene expression sample, NA was assigned to their co-expression value. Genomic distance contains d for those pairs of genes, which are located on different chromosomes.

First gene	**Second gene**	**Co-methylation**	**Co-expression**	**rfunSimAll**	**Genomic distance**
REG1A	REG3A	0.799	0.068	NA	39574
KRTAP13-3	KRTAP15-1	0.799	0.094	NA	15143
SCOC	TNFSF18	0.795	0.029	NA	d
KRTAP21-1	KRTAP13-4	0.795	NA	NA	326231
F7	C1orf64	0.794	0.388	NA	d
KRTAP13-3	KRTAP13-4	0.793	0.051	NA	4897
SPRR3	SPRR1B	0.792	0.502	0.86	30365
C3orf32	TNFSF18	0.789	-0.122	NA	d
CYP11B1	CACNG5	0.789	0.352	0.22	d
PRB4	SCOC	0.788	-0.078	NA	d
SPRR1A	LCE3D	0.787	0.578	NA	403692
SLC5A12	SERPINB12	0.787	0.012	0.01	d
TAS2R13	TBX19	0.787	-0.038	0.21	d
AHSG	FETUB	0.786	0.246	0.71	25674
TAS2R13	C3orf32	0.786	0.083	NA	d
AHSG	CACNG5	0.786	0.231	0.23	d
CDH5	SPARCL1	0.784	0.509	0.61	d
TNFRSF1B	PODN	0.783	0.232	0.56	41150913
SPRR4	LCE5A	0.783	-0.029	NA	459535
KRTAP13-3	KRTAP13-1	0.782	-0.054	NA	29206
TAS2R13	C15orf21	0.782	-0.020	NA	d
SCRG1	SPARCL1	0.782	0.132	NA	85887912
TGIF2LY	TGIF2LX	0.781	NA	0.97	d

Continued

TABLE B.2: The 187 strongest correlations for pairs of genes with respect to their -values, obtained after three stage filtering. For these pairs of genes, we also computed the Pearson correlation of their expression values (fourth column). In case at least one out of two genes was not found in the gene expression sample, NA was assigned to their co-expression value. Genomic distance contains d for those pairs of genes, which are located on different chromosomes.

First gene	**Second gene**	**Co-methylation**	**Co-expression**	**rfunSimAll**	**Genomic distance**
KRTAP13-3	KRTAP20-1	0.781	NA	NA	190684
WNT2	RASGRF2	0.781	0.426	0.46	d
SCOC	C1orf64	0.780	0.241	NA	d
LOR	LCE5A	0.780	0.111	NA	748673
NDRG2	SPARCL1	0.780	0.219	0.55	d
F7	C3orf32	0.779	0.107	NA	d
F7	TNFSF18	0.779	0.024	0.60	d
CRCT1	LCE5A	0.779	0.055	NA	3660
KRTAP8-1	KRTAP13-4	0.779	0.182	NA	382536
DNM3	POU4F1	0.778	0.132	0.48	d
CD8A	SIX6	0.778	NA	0.27	d
REG1B	REG3A	0.777	0.113	0.92	71269
PRB4	SERPINB12	0.777	0.081	0	d
ATP8A2	INA	0.775	0.020	0	d
C1orf64	TNFSF18	0.775	0.051	NA	155083744
APCS	REG1B	0.774	-0.049	0.67	d
ST8SIA5	PODN	0.774	-0.091	0.08	d
LCE3D	LCE2B	0.773	0.191	NA	106303
SERPINB4	SERPINB7	0.773	0.423	0.92	130561
CDX2	INA	0.772	-0.009	0.57	d
COX7B	MAGT1	0.772	NA	0.47	3907
WNT2	POU3F3	0.772	-0.121	0.49	d
SPRR1B	PGLYRP3	0.772	0.309	0.16	279672

Continued

TABLE B.2: The 187 strongest correlations for pairs of genes with respect to their -values, obtained after three stage filtering. For these pairs of genes, we also computed the Pearson correlation of their expression values (fourth column). In case at least one out of two genes was not found in the gene expression sample, NA was assigned to their co-expression value. Genomic distance contains d for those pairs of genes, which are located on different chromosomes.

First gene	**Second gene**	**Co-methylation**	**Co-expression**	**rfunSimAll**	**Genomic distance**
CUBN	SERPINB12	0.771	-0.128	0.47	d
GREM1	PODN	0.771	0.500	NA	d
PRB4	TNFSF18	0.771	0.014	0.33	d
DNM3	SH3TC2	0.771	0.296	NA	d
KRTAP8-1	KRTAP11-1	0.770	0.319	NA	68599
FGF6	CACNG5	0.770	0.202	0.11	d
SLC18A3	PTF1A	0.770	0.000	0.25	26967317
CD8A	POU3F3	0.770	NA	0.31	17966756
APCS	CRP	0.770	0.111	0.77	126545
SIGLEC9	KLK3	0.770	-0.127	NA	269952
NEUROG1	GLB1L3	0.769	-0.154	NA	d
PRB4	C15orf21	0.769	-0.057	NA	d
CD8A	GCM2	0.769	NA	0.21	d
CUBN	SPP2	0.769	0.044	0.36	d
TMEM132D	GJD2	0.768	NA	0.39	d
UBD	OR12D3	0.768	-0.068	0.23	185440
F7	PRR15L	0.767	NA	NA	d
LCE2B	LCE5A	0.766	-0.048	NA	175122
SCOC	C15orf21	0.766	0.195	NA	d
POU4F1	INA	0.766	0.431	0.6	d
SPRR1B	LCE3D	0.766	0.562	NA	451828
PRR15L	TOM1L1	0.766	NA	NA	6943240
NDRG2	CD9	0.765	-0.095	0.55	d

Continued

TABLE B.2: The 187 strongest correlations for pairs of genes with respect to their -values, obtained after three stage filtering. For these pairs of genes, we also computed the Pearson correlation of their expression values (fourth column). In case at least one out of two genes was not found in the gene expression sample, NA was assigned to their co-expression value. Genomic distance contains d for those pairs of genes, which are located on different chromosomes.

First gene	**Second gene**	**Co-methylation**	**Co-expression**	**rfunSimAll**	**Genomic distance**
F7	SCOC	0.765	0.032	NA	d
C3orf32	PRB4	0.765	0.180	NA	d
CD38	PODN	0.765	-0.091	0.41	d
SPRR3	SPRR1A	0.765	0.414	0.92	17771
DNM3	CORIN	0.765	0.053	0.26	d
TMCO5A	SPP2	0.765	NA	NA	d
CDX2	ST8SIA5	0.763	0.093	0.19	d
CD8A	TOX2	0.763	NA	0.19	d
LCE3D	CRCT1	0.763	0.394	NA	65159
PROK2	TBX21	0.762	0.266	0.29	d
FGF6	CYP11B1	0.762	0.187	0.21	d
IL23A	C12orf34	0.761	-0.064	NA	53637741
CD8A	IHH	0.761	NA	0.52	132761597
GREM1	INA	0.760	-0.015	NA	d
TNFRSF1B	IL23A	0.760	0.229	0.54	d
CD8A	POU4F1	0.760	NA	0.24	d
CD8A	INA	0.759	NA	0.31	d
H2AFY	PODN	0.759	-0.035	0.16	d
CD34	PDCD1LG2	0.759	0.028	0.52	d
CD8A	SOX8	0.759	NA	0.37	d
SCOC	TBX19	0.758	-0.167	NA	d
FLG	CRCT1	0.758	0.247	NA	188257
SPRR4	LOR	0.758	-0.017	NA	289138

Continued

TABLE B.2: The 187 strongest correlations for pairs of genes with respect to their -values, obtained after three stage filtering. For these pairs of genes, we also computed the Pearson correlation of their expression values (fourth column). In case at least one out of two genes was not found in the gene expression sample, NA was assigned to their co-expression value. Genomic distance contains d for those pairs of genes, which are located on different chromosomes.

First gene	**Second gene**	**Co-methylation**	**Co-expression**	**rfunSimAll**	**Genomic distance**
MSC	ISL1	0.757	0.075	0.82	d
CDH5	SH3TC2	0.757	-0.042	NA	d
FLG	LCE3D	0.756	0.212	NA	253416
REG1A	OR12D3	0.756	0.060	NA	d
CD8A	PROK2	0.756	NA	0.71	d
CD8A	TBX21	0.755	NA	0.36	d
KRTAP13-3	KRTAP8-1	0.755	0.018	NA	387433
CD8A	RASGRF2	0.755	NA	0.46	d
ST8SIA5	INA	0.755	0.096	0.16	d
KRTAP13-3	KRTAP21-1	0.755	NA	NA	331128
SIGLEC9	PCBP3	0.755	0.013	0.08	d
IHH	POU3F3	0.754	-0.094	0.44	114794841
SIX6	INA	0.754	0.413	0.49	d
ADAMTS4	NDUFS2	0.754	0.181	NA	536
MAPK4	AHSG	0.754	0.201	NA	d
PPP3R2	MORC1	0.754	0.278	NA	d
PPP3R2	REG1B	0.754	-0.005	NA	d
CD244	SLAMF1	0.754	0.746	0.65	216690
SLC9A3	NEUROG1	0.754	0.124	0.07	134322082
RECK	POU3F3	0.753	-0.158	NA	d
APCS	CD1E	0.753	-0.037	NA	1233555
TBC1D8B	PSMD10	0.753	0.232	0.49	1288760
KRT1	IQCF2	0.753	NA	NA	d

Continued

TABLE B.2: The 187 strongest correlations for pairs of genes with respect to their -values, obtained after three stage filtering. For these pairs of genes, we also computed the Pearson correlation of their expression values (fourth column). In case at least one out of two genes was not found in the gene expression sample, NA was assigned to their co-expression value. Genomic distance contains d for those pairs of genes, which are located on different chromosomes.

First gene	**Second gene**	**Co-methylation**	**Co-expression**	**rfunSimAll**	**Genomic distance**
TAS2R13	C1orf64	0.753	0.024	NA	d
TMCO5A	CLDN16	0.752	NA	NA	d
CD8A	CLIP4	0.752	NA	NA	57679313
SIGLEC9	AHSG	0.752	-0.020	0.11	d
FGD4	SH3TC2	0.752	0.314	NA	d
REG1B	OR12D3	0.752	0.175	0	d
KRTAP21-1	KRTAP11-1	0.752	NA	NA	124904
KRT1	AHSG	0.751	0.027	0.38	d
APCS	REG1A	0.751	0.092	NA	d
PGLYRP3	SPRR4	0.751	0.009	NA	340784
ST8SIA5	NEUROG1	0.751	0.198	0.17	d
DNM3	PODN	0.751	0.243	0.38	116776395
CD1E	CD1A	0.751	0.589	NA	100397
PPP3R2	GPR1	0.751	-0.086	NA	d
SCN7A	SERPINB12	0.751	-0.012	0.13	d
CD8A	DNM3	0.751	NA	0.35	d
NPY	INA	0.751	0.277	0.43	d
IHH	ST8SIA5	0.751	0.062	0.19	d
ST8SIA5	TBX21	0.751	-0.138	0.17	d
SPRR4	LCE3D	0.751	0.091	NA	390716
SH3TC2	NDRG2	0.750	0.402	NA	d
SPARCL1	C15orf21	-0.752	-0.037	NA	d
TFF1	C12orf34	-0.755	0.130	NA	d

Continued

TABLE B.2: The 187 strongest correlations for pairs of genes with respect to their -values, obtained after three stage filtering. For these pairs of genes, we also computed the Pearson correlation of their expression values (fourth column). In case at least one out of two genes was not found in the gene expression sample, NA was assigned to their co-expression value. Genomic distance contains d for those pairs of genes, which are located on different chromosomes.

First gene	**Second gene**	**Co-methylation**	**Co-expression**	**rfunSimAll**	**Genomic distance**
DNM3	TAS2R13	-0.756	-0.066	0.09	d
C12orf34	TNFSF18	-0.763	0.002	NA	d
TAS2R13	SH3TC2	-0.775	0.023	NA	d