Establishment of predictive blood-based signatures in medical large scale genomic data sets: Development of novel diagnostic tests

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Universität Bonn

vorgelegt von

Andrea Hofmann geb. Gaarz

aus

Ratingen

Bonn, Oktober 2012

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1.Gutachter: Prof. Dr. Joachim L. Schultze
2.Gutachter: Prof. Dr. Holger Fröhlich

Tag der Promotion: 07. Mai 2013
Erscheinungsjahr: 2013

Eidesstattliche Erklärung

Diese Dissertation wurde im Sinne von § 6 der Promotionsordnung vom 03.Juni 2011 im Zeitraum von April 2008 bis August 2012 von Herrn Prof. Dr. Schultze betreut.

Hiermit versichere ich, dass

- die vorgelegte Arbeit – abgesehen von den ausdrücklich bezeichneten Hilfsmitteln – persönlich, selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt wurde,
- die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte unter Angabe der Quelle kenntlich gemacht sind,
- die vorgelegte Arbeit oder ähnliche Arbeiten nicht bereits anderweitig als Dissertation eingereicht worden ist bzw. sind, sowie eine Erklärung über frühere Promotionsversuche und deren Resultate,
- für die inhaltlich-materielle Erstellung der vorgelegten Arbeit keine fremde Hilfe, insbesondere keine entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder andere Personen) in Anspruch genommen wurde sowie keinerlei Dritte vom Doktoranden unmittelbar oder mittelbar geldwerte Leistungen für Tätigkeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Arbeit stehen.

Bonn, den 17. Oktober 2012

....

Andrea Hofmann

Summary

Increasing data has led to tremendous success in discovering molecular biomarkers based on high throughput data. However, the translation of these so-called genomic signatures into clinical practice has been limited. The complexity and volume of genomic profiling requires heightened attention to robust design, methodological details, and avoidance of bias. During this thesis, novel strategies aimed at closing the gap from initially promising pilot studies to the clinical application of novel biomarkers are evaluated.

First, a conventional process for genomic biomarker development comprising feature selection, algorithm and parameter optimization, and performance assessment was established. Using this approach, a RNA-stabilized whole blood diagnostic classifier for non-small cell lung cancer was built in a training set that can be used as a biomarker to discriminate between patients and control samples. Subsequently, this optimized classifier was successfully applied to two independent and blinded validation sets. Extensive permutation analysis using random feature lists supports the specificity of the established transcriptional classifier.

Next, it was demonstrated that a combined approach of clinical trial simulation and adaptive learning strategies can be used to speed up biomarker development. As a model, genome-wide expression data derived from over 4,700 individuals in 37 studies addressing four clinical endpoints were used to assess over 1,800,000 classifiers. In addition to current approaches determining optimal classifiers within a defined study setting, randomized clinical trial simulation unequivocally uncovered the overall variance in the prediction performance of potential disease classifiers to predict the outcome of a large biomarker validation study from a pilot trial. Furthermore, most informative features were identified by feature ranking according to an individual classification performance score.

Applying an adaptive learning strategy based on data extrapolation led to a data-driven prediction of the study size required for larger validation studies based on small pilot trials and an estimate of the expected statistical performance during validation. With these significant improvements, exceedingly robust and clinically applicable gene signatures for the diagnosis and detection of acute myeloid leukemia, active tuberculosis, HIV infection, and non-small cell lung cancer are established which could demonstrate disease-related enrichment of the obtained signatures and phenotype-related feature ranking.

In further research, platform requirements for blood-based biomarker development were exemplarily examined for micro RNA expression profiling. The performance as well as the technical sample handling to provide reliable strategies for platform implementation in clinical applications were investigated.

Overall, all introduced methods improve and accelerate the development of biomarker signatures for molecular diagnostics and can easily be extended to other high throughput data and other disease settings.

Contents

List of Figures

List of Tables

Nomenclature

AIDS	Acquired immunodeficiency syndrome
ALL	Acute lymphoblastic leukemia
ALS	Adaptive learning set
AML	Acute myeloid leukemia
ART	Antiretroviral therapy
ATB	Active tuberculosis
AUC	Area under the ROC curve
CLL	Chronic lymphocytic leukemia
CML	Chronic myeloid leukemia
COPD	Chronic obstructive pulmonary disease
CT	Computed tomography
Ct	Cycle treshold
CV	Cross-validation
DE	Differentially expressed
DNA	Deoxyribonucleic acid
EFS	Event-free survival
FC	Fold-change
FDA	US Food and Drug Administration
FN	False negatives
FNR	False negtative rate
FP	False positives

FPR	False positive rate
GEX	Gene expression profiling
GO	Gene ontology
HIV	Human immunodeficiency virus
HSC	Hematopoietic stem cell
LDA	Linear discriminant analysis
LMW	Low molecular weight
LTNP	Long-term non progressor
MAQC	MicroArray Quality Control
MCC	Matthews correlation coefficient
MIAME	The minimum information about a microarray experiment
miRNA	Micro RNA
MM	Multiple myeloma
mRNA	Messenger RNA
NPV	Negative predictive value
NSC	Nearest shrunken centroid
NSCLC	Non-small cell lung cancer
OS	Overall survival
PAM	Prediction analysis of microarrays
PBMC	Peripheral blood mononuclear cells
PCR	Polymerase chain reaction
PN	Predicted negatives
PP	Predicted positives
PPV	Positive predictive value
PTS	Pilot trial study

QC	Quality control
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
SAM	Sentrix Array Matrix
SCLC	Small cell lung cancer
Sens	Sensitivity
Spec	Specificity
SVM	Support vector machine
TB	Tuberculosis
TN	True negatives
TP	True positives
TPR	True positive rate
TS	Training set
TSA	Trial simulation approach
VS	Validation set
WHO	World Health Organization
YI	Youden Index

Preface

This work was carried out in the Genomics and Immunoregulation group of the Life and Medical Science Institute (LIMES) at the University of Bonn and it is a pleasure to thank the many people who supported me during tough times and made this thesis possible. First and foremost, I am very grateful to Prof. Dr. Joachim L. Schultze for giving me the opportunity to write my thesis under his guidance. His enthusiasm, continuous support and inspiring discussions contributed most to the success of this project and it is a pleasure to thank him for all his time, brilliant ideas, encouragement, advice, funding and necessary freedom to pursue my scientific work. I would like to thank Prof. Dr. Holger Fröhlich for his willingness to be the co-referent of the thesis and PD Dr. Gerhild van Echten-Deckert and Prof. Dr. Rainer Manthey for their participation in the committee.

I would also like to thank all past and present colleagues for the excellent working atmosphere and the fruitful discussions especially PD Dr. Andrea Staratschek-Jox with whom I worked in close collaboration on almost every project. I gratefully acknowledge Dr. Svenja Pascher, Dr. Sabine Classen, Dr. Marc Beyer, Dr. Michael Mallmann and Fatima Kreusch for their contributions to support and improve my work. It has been a pleasure to supervise the graduate students Melanie Henseler, Saphira Wollmer and Sibel Gördüm and the students Benjamin Engelhardt, Arne Schenk and Thileepan Sekeran assisted with data collection and analysis. Special thanks goes to Michael Kraut for outstanding technical quality of data and to our system administrator Tom Wegner for his sound help in different computer related matters. I further thank Thomas Ulas for his expert corrections of this thesis and intensive re-analysis of all my scripts and implementations.

I owe a huge debt of gratitude to the LIMES graduate school for providing me a scholarship to finish this research.

Finally, I would like to thank Prof. Dr. Maik Kschischo and David Endesfelder from the RheinAhrCampus, Prof. Dr. Holger Fröhlich from the B-IT, Prof. Dr. Rudy Parrish from the University of Louisville and Dr. Thomas Zander from the University Hospital Cologne for good collaborations. Their considerable expertise and knowledge were invaluable contributions to this thesis.

And a more private note, I would like to thank my family and all people contributed to child care and most importantly; special thanks to my loving, supportive and encouraging husband Tobias Hofmann for standing by me through all these years.

I dedicate this thesis to my beautiful children Mathilda and Theo.

The results presented here were obtained in close collaboration with many different people and parts have been published before. I would therefore like to elaborate my specific contribution to each of the projects.

Chapter 4 summarizes the development of blood-based gene expression signatures for the identification of lung cancer. The research on this topic was performed in close collaboration with Dr. Thomas Zander and PD Dr. Andrea Staratschek-Jox. The bioinformatic parts of the study including statistical analysis were carried out by me. The results of this study have been published in *Clinical Cancer Research* (Zander, Hofmann et al, 2011).

Chapter 5 introduces methods to enhance molecular diagnostics applied to various gene expression data sets. Microarray experiments and sample collection were performed by PD Dr. Andrea Staratschek-Jox, Dr. Svenja Debey-Pascher, Michael Kraut and Mirela Stecki in collaboration with the University Hospital Cologne. Original analysis strategies were developed by Prof. Dr. Joachim L. Schultze, PD Dr. Andrea Staratschek-Jox and me and all bioinformatic procedures including data collection were inplemented by me. This work was further extended and substancially improved in close discussion with Prof. Dr. Maik Kschischo, David Endesfelder, Prof. Dr. Holger Fröhlich and Prof. Dr. Rudy Parrish. The oversampling strategy was developed and carried out by Prof. Dr. Joachim L. Schultze. The results of this study are currently prepared as a manuscript for submission and parts of the study have been presented as a poster at the ISMB/ECCB 2009 in Stockholm. All bioinformatic scripts have been tested and verified by Thomas Ulas.

In chapter 6, a microRNA microarray platform is evaluated for blood expression profiling. Experiments were performed by Dr. Svenja Debey-Pascher and Michael Kraut; all bioinformatic analysis was carried out by me. PD Dr. Andrea Staratschek-Jox supervised this project and results have been published in the *Journal of Molecular Diagnostics* (Gaarz, Debey-Pascher et al, 2010).

This thesis reproduces figures from our publications. Figures Fig. 4.2.1, Fig. 4.2.2, Fig. 6.2.1, Fig. 6.2.2 and Fig. 6.2.3 are reproduced with permission from these journals.

Andrea Hofmann — Bonn, October 2012

1. Introduction

The introduction of genome-wide gene expression profiling by microarrays in the mid-1990s (*Schena et al.*, 1995) enabled multidimensional measurement of biological processes in parallel. Researchers investigated in the enormous potential of this technology for both gaining insights into molecular biology as well as its use in clinically motivated applications. Golub et al. were the first to successfully demonstrate microarray-based transcriptional classification to distinct acute myeloid and acute lymphocytic leukemia (*Golub et al.*, 1999). Further landmark studies in cancer research have highlighted the power of this technology to classify types of tumors (*Alizadeh et al.*, 2000; *Bittner et al.*, 2000; *Sorlie et al.*, 2001; *Khan et al.*, 2001; *Yeoh et al.*, 2002) and to predict the outcome (*Beer et al.*, 2002; *van de Vijver et al.*, 2002; *van 't Veer et al.*, 2002; *Pomeroy et al.*, 2002; *Shipp et al.*, 2002; *Bullinger et al.*, 2004; *Valk et al.*, 2004) and even the response to chemotherapy (*Kihara et al.*, 2001; *Rosenwald et al.*, 2002). A major challenge with high-throughput technologies is the analysis of a complex data output. Different approaches and mathematical algorithms were utilized to classify patients on the basis of expression profiles (*Yeang et al.*, 2001; *Ramaswamy et al.*, 2001; *Dudoit et al.*, 2002a; *Simon*, 2003).

Nevertheless, despite a growing number of studies on transcriptional-based biomarkers, only a handful of gene signatures have entered into clinical practice to date. One prominent example is the prognostic breast cancer signature MammaPrint®. In 2002, scientists established a 70-gene signature for the identification of lymph node negative breast cancer patients at high risk for disease progression (*van 't Veer et al.*, 2002) and further validated this test to assess prognosis of distant metastasis in a series of independent studies (*van de Vijver et al.*, 2002; *Glas et al.*, 2006; *Bueno-de Mesquita et al.*, 2007). In 2007, the US Food and Drug Administration (FDA) proved the MammaPrint and further large multi-institutional clinical trials are assessing this test (*Cardoso et al.*, 2008). Still, in comparison to thousands of studies reporting microarray-based classifiers, the resulting number of clinical test is deficient.

Furthermore, the initial enthusiasm was tempered by serious concerns on the robustness and reproduciblity of the methodology (*Ioannidis*, 2005; *Liu and Karuturi*, 2004; *Michiels et al.*, 2005). Microarray technology is particularly susceptible, because measurement errors can occure from sampling, preprocessing, processing, calibration and data analysis. Focussing on technical aspects, little overlap was reported among interplatform and cross-laboratory comparison studies (*Ramalho-Santos et al.*, 2002; *Tan et al.*, 2003; *Miller et al.*, 2004). Already in 2001, researchers need to present accurate information on crucial aspects of measurements let to the use of highly standardized protocols documented in the minimum information about a microarray experiment (MIAME) guidelines for reporting, annotation, and data analysis of microarray data (*Brazma et al.*, 2001). As an important result of concerns about the reliability of the technology, the MicroArray Quality Control (MAQC)-I project systematically addressed the impact of different microarray platforms and lab-to-lab variability in reproducibility and comparability of microarray results. With the aim to provide quality control tools to the microarray community to avoid procedural failures the study successfully demostrated that the technology itself is reliable and reproducible (*Shi et al.*, 2006; *Guo et al.*, 2006; *Kuo et al.*, 2006; *Canales et al.*, 2006; *Shippy et al.*, 2006).

Despite the clear demonstration of microarray robustness performed under stringent conditions, further doubts have been raised by non-overlapping gene signatures derived by several studies adressing the same clinical outcome (*Dupuy and Simon*, 2007; *Sotiriou and Piccart*, 2007) questioning their biological significance and clinical implications. However, for example in breast cancer profiling most classifiers although having only few genes in common show a significant concordance in outcome prediction (*Fan et al.*, 2006; *Ein-Dor et al.*, 2005). Another major drawback of microarray studies are small sample sizes in the original studies and a lack of independent validation sets leading to overoptimistic claims that can arise from uncommon, fragmented and incomplete validation (*Simon et al.*, 2003; *Ntzani and Ioannidis*, 2003). Michiels et al. employed a multiple random validation strategy on published studies to show that because of inadequate validation, originally reported assessments are overoptimistic and prognostic values should be considered with caution (*Michiels et al.*, 2005).

Recently, these concerns about biased results could be eased by improving the analytical processes used within these pilot studies (*Fan et al.*, 2010a). In addition, the second phase of the MAQC project (MAQC-II) set the framework for making clinically useful predictions from large-scale gene expression data (*Shi et al.*, 2010; *Oberthuer et al.*, 2010; *Luo et al.*, 2010; *Parry et al.*, 2010; *Huang et al.*, 2010; *Fan et al.*, 2010b). Several important issues are solved by this large consortium effort sought to evaluate various data analysis methods in developing and validating microarray-based predictive models. Most important, model prediction performance was found to depend heavily on the endpoint, probably the most critical finding supporting further development of gene signature technology. Further, internal validation performance from well-implemented, unbiased cross-validation shows a high degree of concordance with blinded external validation performance. Nevertheless, external validation is a critical feature for signature development. Formerly questioned by others, the analysis also clearly established that many classifiers with similar performance can be developed from a given data set. Not surprising, proficiency of investigators and good modeling practice are leading to improved results. Overall by collecting and analyzing more than 18,000 predictive models, the MAQC-II highlighted that rigorous standards for reporting the analytical steps are neccessary and systematically gave evidence that predictive models can be reliable enough to justify clinical applications.

These comprehensive findings strongly suggest to bring transcriptional-based models into clinical practice. Since a key challenge in cancer medicine is to detect a disease as early as possible to improve treatment and reduce the mortality, the establishement of simple, fast and robust clinical diagnostic biomarkers are highly desirable. For many disease states, the primary affected tissues are not readily available or can only be obtained by invasive intervention. Alternativly, the use of the surrogate tissue peripheral blood for transcriptional profiling may be feasible for the development of new diagnostic markers since it is the most accessible and practical source of messenger RNA (mRNA) in patients and is in contact with almost every organ and tissue in the body (*Baird*, 2006; *Burczynski and Dorner*, 2006). Based on the assumption that circulating blood might reflect pathological changes occuring in different tissues of the body (*Liew et al.*, 2006), an increasing number of clinical studies monitored blood-based biomarkers by gene expression profiling (*Twine et al.*, 2003; *Burczynski et al.*, 2005; *Sharma et al.*, 2005; *Osman et al.*, 2006; *Critchley-Thorne et al.*, 2007; *Showe et al.*, 2009; *Staratschek-Jox et al.*, 2009; *Chaussabel et al.*, 2010).

Within this thesis, methods are established to improve and accelerate the development of transcriptional blood-based diagnostic biomarkers from explorative pilot studies to their clinical routine use. The complexity and volume of transcriptional profiling requires heightened attention to robust design, methodological details and avoidance of bias. It is structured as follows:

In *chapter 2*, I will present all background knowledge that is needed to understand and follow this thesis. The fundamentals of molecular biology will be explained as well as genome-wide transcriptional profiling using microarrays followed by sketching basic microarray data analysis methods. This section mainly focusses on microarray classification analysis including study design, major classification algorithms and performance measurements.

Chapter 3 briefly introduces the most important Material and Methods used for this thesis covering data sets used and bioinformatic methods applied. Further information on the experimental setups can be found in (*Zander et al.*, 2011; *Gaarz et al.*, 2010) and in Appendix C.

In *chapter 4*, a conventional microarray classification workflow is established for the identification of non-small cell lung cancer (NSCLC) in comparison to control samples. This workflow comprises (i) comparison of feature selection cut-offs and classification algorithms in a training set, (ii) building an optimized classifier in the training set using the choosen features and algorithm and (iii) application of this classifier to two independent validation cohorts and assessment of classifier performance. This approach is further applied to one of the MAQC-II gene expression datasets.

Chapter 5 introduces methods to enhance molecular diagnostics including novel strategies to predict the outcome of a large biomarker study from a small pilot study and to estimate the sample size of future pivotal validation cohorts. These concepts are developed on a large acute myeloid leukemia (AML) data set and then applied to molecular diagnosis using peripheral blood of NSCLC, active tuberculosis (ATB) and human immunodeficiency virus (HIV).

In *chapter 6*, I would like to focus on the challenges when a novel high-throughput technology is applied to the development of blood-based diagnostic biomarkers. Reasonable concerns would arise from the reproducibility of expression profiles and the impact of RNA isolation methods. The following study interrogates the latter concern using a recently introduced bead-array based microRNA (miRNA) expression technology.

The last chapter of the thesis gives a summary of all previous discussions and aligns these into a broader context. Here future research directions are pointed out and possible study limitations are discussed.

2. Background

All background knowledge that is needed to understand and follow this thesis as well as biological, technical and bioinformatics terms will be explained in this chapter. Starting with the fundamentals of molecular biology, the definition of the three macromolecules Deoxyribonucleic acid (DNA), Ribonucleic acid (RNA) and proteins will be clarified. Next, genome-wide transcriptional profiling using microarray technology is introduced followed by sketching basic microarray data analysis methods. The following section focuses on microarray classification analysis including study design, major classification algorithms and performance measurements. Finally, all diseases addressed in this thesis will be briefly introduced.

2.1. Molecular biology

The discovery of the DNA's double helix structure by Watson and Crick in 1953 introduced a new view on biology as it raised the question about how biological information is encoded in DNA (*Watson and Crick*, 1953). Remarkable, DNA structure can store complex information in a digital code. Its information covers two features that are essential for all known forms of life: The genes that encode proteins which are involved in regulatory processes of cells and the gene networks that regulate the function of genes (*Hood and Galas*, 2003).

DNA is a nucleic acid containing the genetic information used in the development and functioning of all known living organisms and some viruses. The DNA molecule is composed of so called nucleotides, whereas each nucleotide consists of a desoxyribose, one phosphoric acid and one out of the four organic bases adenine (A), guanine (G), cytosine (C) and thymine (T). The nucleotides are arranged in two long polymer strings which run in opposite directions to each other and are therefore anti-parallel. The double stranded DNA forms a helical structure (*Watson and Crick*, 1953). The four bases of the DNA molecules carry genetic information whereas their sugar and phosphate groups perform a structural role. Almost any sequence of bases and hence, any digital information can be accommodated by the DNA molecule.

In a process called transcription, the DNA is copied into a template for protein synthesis, the RNA molecule. Each gene encodes a complementary RNA transcript called messenger RNA (mRNA) (*Brenner, Jacob, and Meselson*, 1961). Like DNA, RNA consists of nucleotides but the sugar unit in RNA is ribose and the base thymine is replaced by the derivate uracil (U). Additionally, RNA usually is single-stranded, except for some viruses. Additionally to the coding mRNA which serves as a template for protein synthesis, different types of non-coding RNA exist. The most prominent representatives of non-coding RNAs are transfer RNAs (tRNA) and ribosomal RNAs (rRNA). Other non-coding RNAs include antisense RNA (aRNA), microRNA (miRNA) and small interfering RNA (siRNA) which all function as gene regulation molecules.

In the next step in protein synthesis, the translation, a mRNA produced by transcription encodes for the production of a specific amino acid chain by the ribosome that is fold into an active protein (Fig. 2.1.1). Sequences of the four bases A, C, G and T/U of genes are used to decode proteins. Each triplet of nucleotides in a nucleic acid sequence, so called codons, are mapped to one of the 20 amino acids of the protein alphabet (*Crick et al.*, 1961), this information definition is called the genetic code (*Nirenberg and Matthaei*, 1961).

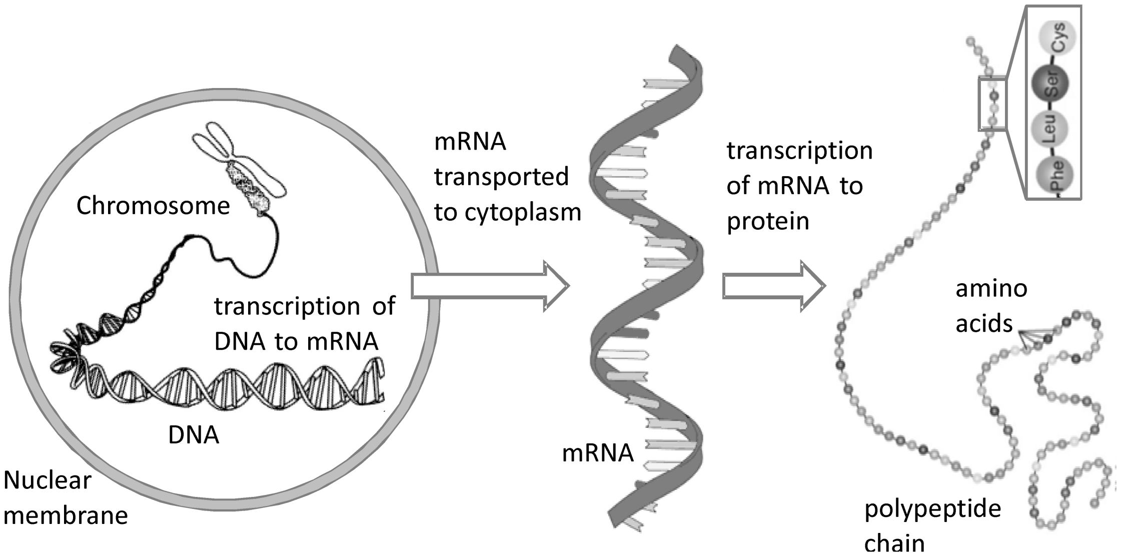

Chromosome
transcription of DNA to mRNA
DNA
Nuclear membrane
mRNA transported to cytoplasm
mRNA
transcription of mRNA to protein
Phe
Leu
Ser
Cys
amino acids
polypeptide chain

Figure 2.1.1.: From DNA to RNA to Proteins:
Process of information transfer. The graphic was modified from *Chisholm et al.* (2007).

Proteins are essentially for all biological processes, most known is the role of proteins as enzymes to catalyze biochemical reactions (*Gutteridge and Thornton*, 2005). Furthermore they transfer cell signals and are involved in signal transduction, they serve as antibodies in the immune system and transport other molecules. Structurally, proteins are build from amino acid chains that are fold together to unique structures by the linkage of peptide bonds.

2.2. Microarray technology

Microarrays are the major technology to assess genome-wide transcriptional profiles of cells, tissues or even whole organs that allow to probe of the expression of thousands of genes simultaneously (*Quackenbush*, 2006b; *Schena et al.*, 1995). A microarray is an orderly arrangement of probes in a 'micro' format in which many objects share a relatively small area. In this thesis, the term 'probe' is used to describe the sequence of nucleotides that are immobilized on the array. The term 'sample' is used to define the mRNA extracted from the biological samples analyzed in a microarray experiment.

The basic principle works as follows (Fig. 2.2.1): A large number of probes containing short nucleic sequences representing thousands of individual genes or other DNA elements are immobilized on a solid surface. RNA is extracted from samples of interest and labeled with a fluorescent or radioactive marker. The pool of labeled mRNA-representing nucleotides, refered to as targets, is hybridized to the probes. Hybridization is the process where complementary strands of nucleotides, in this case those of probe and target, bind to each other by building hydrogen bridges between the complement base pairs (Watson-Crick base pairs). After the hybridization process, probe-target hybridization is detected by measuring the fluorphore-labeled target intensity which determines the relative abundance of nucleic acid sequences in the target and later on is quantified into numerical values. Higher hybridization degrees result in increased signal intensity implying a higher relative level of expression of the particular gene.

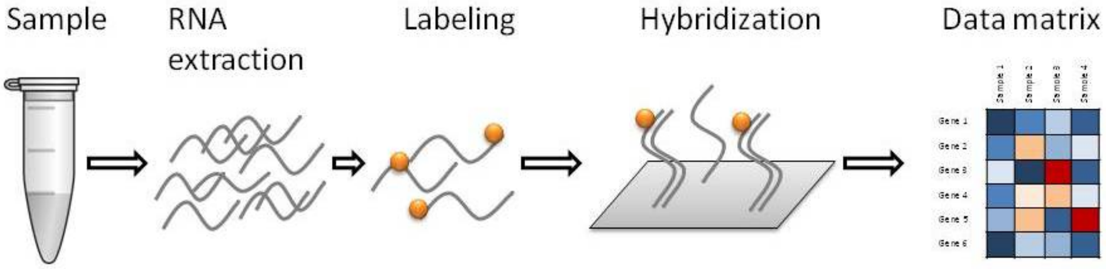


Figure 2.2.1.: Principle of the microarray technology.
RNA is extracted from a sample of interest, labeled with a marker dye and hybridized to complementary gene-specific probes on the array. Relative levels of gene expression is estimated by measuring the fluorescence intensity for each probe and summarized in a expression data matrix. The figure was modified from *Quackenbush* (2006b).

Among the most widely used technologies is the GeneChip distributed by Affymetrix (*Lockhart et al.*, 1996). The Affymetrix GeneChips are a constructed using a combination of two techniques, photolithography and solid-phase DNA synthesis. Probes on these arrays are synthesized using a light mask technology to sequentially build nucleotide sequences. An Affymetrix chip comprise of a number of probesets, each probeset consist of 25mer probe pairs (25 bases) selected from the target sequence: one perfect match and one mismatch for each chosen target position. The mismatch probe contains the same 25 base long sequence as the perfect match probe, except the middle base is substituted for its complement base. Other array types such as those from Illumina are based on randomly assembled arrays of microscopic beads, each with a specific address sequence, which do not interfere with the fluorescent dyes used on the target sequence (*Kuhn et al.*, 2004). Further distributors of DNA microarrays include GE Healthcare, Applied Biosystems, Beckman Coulter, Eppendorf Biochip Systems and Agilent.

2.3. Basic microarray data analysis methods

2.3.1. Quality control

After collection, the data is usually controlled for quality. Quality control (QC) is an important step before further downstream data analysis as it quantifies the measurement quality for any particular sample and hence, identifies outlier samples and detects poor quality or uninformative data. Several methods have been proposed, some of which are depicted here:

Most methods are based on visual inspection of the data, where either the distribution of all samples in the experiment is shown (density or box plots, Fig. 2.3.1A) or pairwise comparisons such as scatterplots (Fig. 2.3.1B) or MA plots (introduced by *Dudoit et al.* (2002b), Fig. 2.3.1C) are performed.

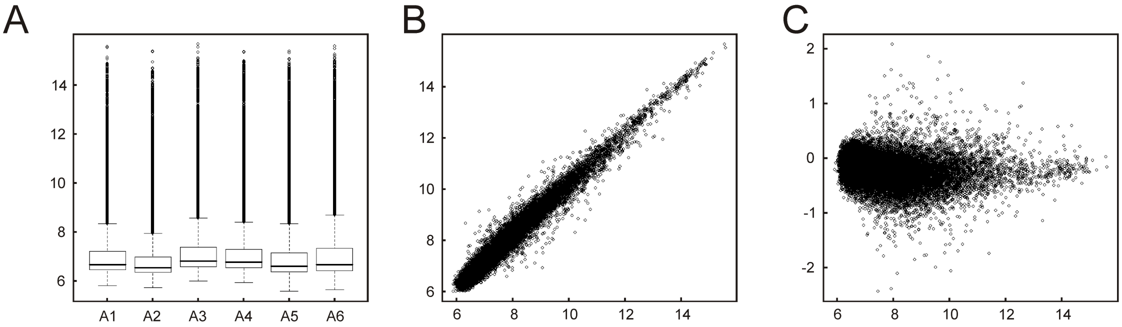


Figure 2.3.1.: Diagnostic plots for microarray quality control
Depicted here are (A) a boxplot, (B) a pairwise scatter plot and (C) a MA plot of log-intensity values.

Box plots (also known as box-and-whisker diagram) present the distribution of signal intensities across all samples of a given data set by summarizing key descriptive statistics. The plot consists of boxes with a central line and two tails. Each box represents the inter-quartile range (IQR) of that sample where the middle 50% of the ranked data are found, whereas the central line state the median of the data. The upper (75% percentile) quartile Q1 and the lower (25% percentile) quartile Q3 are the bottom and top edges of the box, respectively. The tails of the whiskers can represent several values, including minimum and maximum values or 5^{th} and 95^{th} percentiles.

Similarily, the spatial distribution can be visualized by **density plots** to access the homogeneity between the samples of an experiment. Density estimation is the estimation of an underlying probability density function based on observed data (*Silverman*, 1986). Specifying this density function f allows probabilities associated with a random quantity X to be found from the equation

$$P(a < X < b) = \int_{a}^{b} f(d)dx \;\; for\, all\, a < b \tag{2.3.1}$$

Density estimations provide an easy explanation and illustration of data. The total area under the density function integrates to 1.

For pairwise comparisons, for example when comparing biological replicates, a **scatterplot** is one of the simplest methods to visualize expression levels. Each transcript's intensity from one sample is plotted against the intensity value of the other sample and transcripts with similar expression levels in two experiments will appear around the first diagonal of the coordinate system. The correlation coefficiant between both samples can be calculated as well.

A **MA plot** is a rotation of the scatterplot by 45° with a subsequent re-scaling of the data. Both sets of signals are not treated separately. Hence, a transcripts's M-value as the log-ratio of the two intensities is plotted against the A-value as the mean of their logarithms. As most transcripts are expected to have equal expression in both experiments, the majority of points will be grouped around the horizontal line.

Another way to examine the quality of microarrays is checking background hybridization levels by determining the absent and present status of probes (**absent/ present calls**). This is performed by calculating a detection p-value which statistically compares the expression signal to a background control usually present on the microarray. A probe is called present if the expression signal significantly differs from the negative signal, otherwise absent. As a quality control measure, the percentages of present genes indicating the sensitivity of arrays within an experiment should be similar for each array within an experiment.

If any of the quality measurements (different diagnostic plots or percentages of present probes) indicates an obvious strong technical outlier in the data set, the affected sample is usually removed from further analysis.

2.3.2. Normalization

Normalization is an essential procedure that is directed at resolving the systematic errors and bias, introduced by the experimental process in the microarray technology. Different sources of systematic variation such as unequal quantities of starting material, efficiency of RNA extraction, reverse transcription, labeling, photodetection etc. can affect the measured gene expression. Therefore data should be corrected for those effects in order to detect biological differences between RNA samples. Normalization comprises the steps background correction, data transformation (usually on logarithmic scale to deliver approximately normal symmetric distribution allowing the use of a powerful group of statistical tests) and estimation of a "rescaling" factor (using internal controls).

Generally there are two groups of normalization methods to account for systematic errors: The within array normalization to handle error effects on one array individually and the between array normalization to achieve consistency between arrays (*Smyth and Speed*, 2003; *Yang et al.*, 2002a).

The within array normalization method works by subtracting the fitted values on a linear or non-linear regression in one or two dimensions. We obtain the normalized expression value $log_2(p_{ik})$ by

$$log_2(p_{ik}) = log_2(\tilde{p_{ik}}) - f\left(log_2\left(\tilde{p_{ik}}\right)\right) \tag{2.3.2}$$

where $\tilde{p_{ik}}$ is the original expression value of gene i on array k and f is a regression function.

The between array normalization methods are mainly performed by scaling and centering of the data. Most methods are based on the assumption that between two conditions the majority of genes are not supposed to be differentially expressed and only a small group is supposed to be differentially expressed.

Most common methods are

- Scale normalization: Here the scale of the data is simply adjusted to have the same median-absolute-deviation across arrays, e.g. by setting the median of differences to zero (*Yang et al.*, 2002b).
- Quantile: Each of the array-specific distributions of intensities are transformed so that all have the same values at specified quantiles in order to ensure that the intensities have the same empirical distribution across arrays and across channels (*Bolstad et al.*, 2002).
- Variance stabilizing transformation (VSN): The VSN method builds upon the fact that the variance of microarray data depends on the expression intensity and transforms the data to have equal variance for all intensities. It is similar to the natural log transformation at the upper end of the intensity scale but

tries to adjust high-variance effects at the lower end, e.g. from background correction (*Huber et al.*, 2002).

- MAS5: For Affymetrix arrays, additional normalization is necessary using the perfect and mismatch probe sets on the array. MAS5 normalizes each array independently and sequentially by using the mismatch probes to calculate a "robust average", based on subtracting the mismatch probe value from the match probe value (*Hubbell et al.*, 2002).
- Robust multichip average (RMA): The RMA method calculates probe level summaries with empirically motivated statistical models and does not use the mismatch probes due to the fact that their intensities are sometimes higher than the match probes, making them unreliable as indicators of non-specific binding (*Irizarry et al.*, 2003).

2.4. Classification of high-dimensional microarray data

Genomic biomarkers are developed by classification analysis of high-throughput expression patterns. This classification analysis is a supervised learning process in which class memberships are predefined. A predictive model based on a set of labeled observations (the training set) with a set of features is used to build a classifier. The resulting mathematical rules from this classifier can be applied to predict the classes of new observations (*Hastie et al.*, 2001). The process of biomarker development involves three steps:

- identification of informative features
- selection of classification algorithms and model building
- performance assessment.

A general overview of the process to build a predictive model for sample classification is shown in Fig. 2.4.1.

2.4.1. Feature selection

Microarray data simultaneously measures the expression levels of several thousands of transcripts on a small number of samples. Settings in which the number of features p are much larger than the number of observations, in this case samples N ($p \gg N$) can lead to high variance and overfitting problems as the misclassification rate in the training set decrease, but the misclassification rate of new samples might begin to increase (*Hastie et al.*, 2001). Additionally, expression levels can be highly correlated.

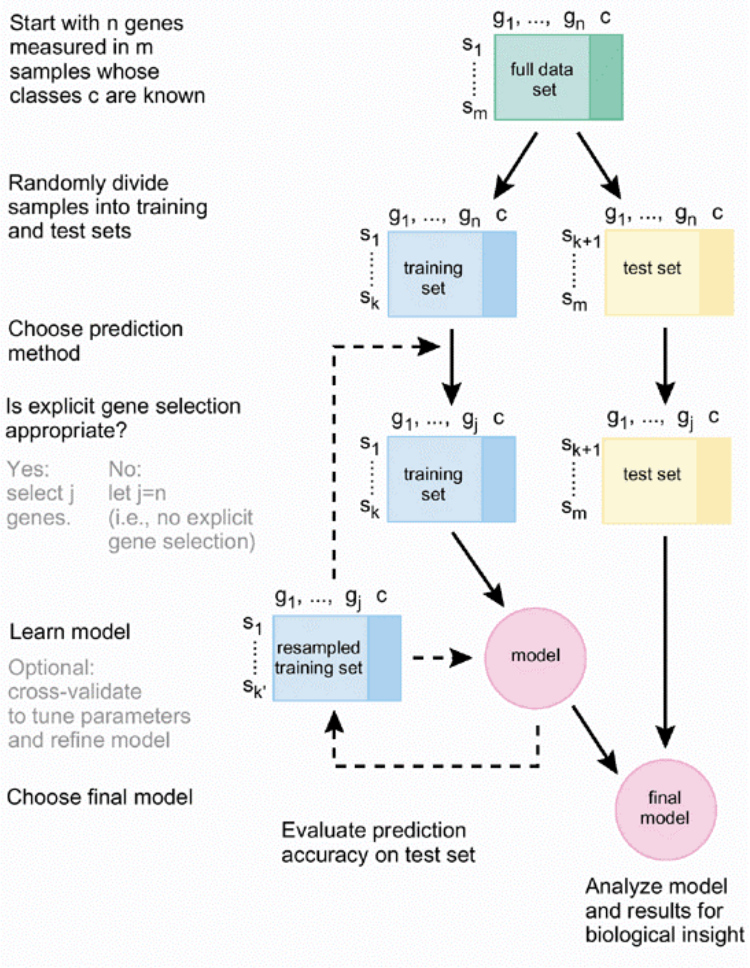

Start with n genes measured in m samples whose classes c are known
$g_1, \dots, g_n$ c
s_1
s_m
full data set
Randomly divide samples into training and test sets
$g_1, \dots, g_n$ c
s_1
s_k
training set
$g_1, \dots, g_n$ c
s_{k+1}
s_m
test set
Choose prediction method
Is explicit gene selection appropriate?
Yes: select j genes.
No: let j=n (i.e., no explicit gene selection)
$g_1, \dots, g_j$ c
s_1
s_k
training set
$g_1, \dots, g_j$ c
s_{k+1}
s_m
test set
Learn model
Optional: cross-validate to tune parameters and refine model
$g_1, \dots, g_j$ c
s_1
$s_{k'}$
resampled training set
model
Choose final model
Evaluate prediction accuracy on test set
final model
Analyze model and results for biological insight

Figure 2.4.1.: An overview of the process for building a prediction model to classify samples
The partition into training and test data is ideally chosen at random across the entire set of samples. Many prediction methods require tuning some parameter such as the number of genes or the number of nearest-neighbors to consider. This choice is often evaluated by cross-validation. The final model is then tested on entirely new data not used in the model generation process. The model itself, as well as the prediction results and the influential genes, may yield new biological insights. Figure is taken from *Slonim* (2002) with permission from the journal.

One common approach to overcome this problem involves the application of feature selection. When eliminating for example transcripts with minimal variance across sample collection the complexity of the dataset is reduced. Hence only informative features are kept showing significantly different signal intensities between two or more groups, often referred to as differentially expressed genes. The most simple method is to calculate the fold-change (FC) between two groups:

$$FC = \frac{m_1}{m_2} \tag{2.4.1}$$

where m_1 and m_2 are the mean expression levels of the two experimental groups. Features with a FC greater then a predefined threshold (e.g. features with at least a 2-fold change) are selected.

However, the FC measure alone is not considered as an adequate test statistic because it does not provide a significance estimate for the observed changes, the thresholds for the FC are arbitrary and it does not incorporate variance (*Allison et al.*, 2006). This problem can be overcome when restating the biological question of differential expression to a problem in statistical hypothesis testing. The intent of statistical hypothesis testing is to determine whether observations made are attributed to chance or provide enough evidence to reject a proposition. The usual process consists of three steps (*Fisher*, 1935; *Lehmann and Romano*, 2005):

1. A conjecture or hypothesis associated with the process, referred to as the **null hypothesis** H_0 (often, that the observations are obtained by chance) and an **alternative hypothesis** H_1 are formulated. Importantly, H_0 and H_1 reference population values and not observed statistics. In this case the null hypothesis can be claimed as no differential expression between the experimental groups.

2. A **test statistic** that measures deviations from the H_0 is identified to test the hypothesis. The statistical testing method must be chosen depending on the underlying data, for instance, Welch t-statistics, Wilcoxon statistics, F-statistics or paired t-statistics can be used for testing. The test statistic from experimental observations is compared to its know sampling distributions.

3. The corresponding **p-value**, the associated probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming H_0 is true is determined. The p-value is then compared to an acceptable significance value α and H_0 is rejected if the p-value is equal or less than α (p-value$\leq \alpha$), otherwise H_0 is retained.

The Welch t-test is the most commonly applied test in micoarray testing and defined as follows:

$H_0 : \mu_X = \mu_Y$
$H_1 : \mu_X \neq \mu_Y$

with samples X and Y exhibiting unequal variance ($\sigma_X \neq \sigma_Y$) and normal distribution ($X \sim N(\mu_X, \sigma_X^2); Y \sim N(\mu_Y, \sigma_Y^2)$. The test statistic is defined as

$$T(X,Y) = \frac{|\overline{X} - \overline{Y}|}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}} \tag{2.4.2}$$

with $\overline{X}$ and$\overline{Y}$ as the means of X and Y, s_X^2 and s_Y^2 as the estimated variances of X and Y and n_1 and n_2 as the sample sizes of X and Y, respectively. The degrees of freedom ν can be calculated as

$$\nu = \left(\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}\right)^2 / \left(\frac{(s_X^2/n_1)^2}{n_1 - 1} + \frac{(s_Y^2/n_2)^2}{n_2 - 1}\right) \tag{2.4.3}$$

The decision to reject H_0 is done on a significance level α if $|T| \geq t_{v;1-\frac{\alpha}{2}}$.

All hypothesis tests implicate the risk for making the wrong conclusion. In any testing situation, independent from the test statistic used, two types of errors can be committed (see Tab. 2.1).

		truth	
		H_0 is true	H_0 is wrong
decision	H_0 accepted	✓	type II error
	H_0 rejected	type I error	✓

Table 2.1.: Types of errors in hypothesis testing
Two types of errors can be committed in a testing situation: A type I error, which is the wrong decision that is made when a test rejects a true null hypothesis H_0 and a type II error, which is the wrong decision that is made when a test fails to reject a false null hypothesis H_0.

A type I error, is defined as the rejection of a true null hypothesis H_0. Its rate α is the probability of a type I error and is also called the significance level of a test. Typical choices for α are 0.05 or 0.01. A type II error, is defined as the failure of rejecting a false null hypothesis and its probability is denoted by the symbol β. Statistical tests can be compared by their power $(1 - \beta)$ to detect true positives. For the analysis of differentially expression, α is committed by declaring a feature to be differentially expressed when it is not, and β is committed when the test fails to identify a truly differentially expressed feature (*Allison et al.*, 2006).

2.4.2. Multiple testing corrections

All the different statistical approaches to find the differentially expressed features suffer from the problem of multiple testing since thousands of genes are analyzed simultaneously and the number of features is much larger than number of samples ($p \gg N$) (*Quackenbush*, 2006a), for example if 10,000 features are tested at a significance level $\alpha = 0.05$, a total of 5% or 500 features might be called significant by chance alone. This leads to hundreds of false positives by chance when tens of thousands of features are tested. To compensate the error of multiple testing, a correction must be done. Most frequently used methods are listed in Tab. 2.2.

Method	Type of error control	Stringency
Bonferroni	FWER	more false negatives
Holm-Bonferroni	FWER	
Westfall and Young Permutation	FWER	↕
Benjamini and Hochberg	FDR	
None	none	more false positives

Table 2.2.: Methods for multiple testing correction
The methods are listed in order of their stringency, with the Bonferroni correction being the most stringent, and the Benjamini and Hochberg FDR being the least stringent. The more stringent a multiple testing correction, the less false positive features are identified. The trade-off of a stringent multiple testing correction is that the rate of false negatives is very high.

All methods control either the family-wise error rate (FWER) or the false discovery rate (FDR). The FWER is defined as the probability that at least one false positive error will be committed (*Dudoit et al.*, 2002b) and the FDR is the expected proportion of Type I errors among the rejected hypotheses.

In the **Bonferroni correction** (*Bonferroni*, 1936) the p-value is multiplied with the number of tests performed ($\widetilde{p} = min[p{*}n_{feat}, 1]$; n_{feat} number of features). If the adjusted p-value $\widetilde{p}$ is still below the error rate, the feature will be called significant. This method is very conservative and leads to a high number of false positives.

A less stringent method is the **Holm-Bonferroni Step-down correction** (*Holm*, 1979). The p-value of all features are ranked with $p_1 \leq ... \leq p_k \leq ... \leq p_{n_{feat}}$and then is $\widetilde{p_j} = \underset{k=1,\ldots,n_{feat}}{max} \{min[p_k * (n_{feat} - k + 1), 1]\}$. Both methods are called single-step procedures, where each p-value is corrected independently.

In contrast to this, the **Westfall and Young permutation method** (*Westfall and Young*, 1993) takes advantage of the dependence structure between features and is based on a resampling procedure to compute p-values. The proportion of resampled data sets where the minimum resampling p-value is less than the original p-value is the adjusted p-value. More information can be obtained in (*Ge et al.*, 2003).

Rather than controlling the probability under the null hypothesis, the **Benjamini-Hochberg method** (*Benjamini and Hochberg*, 1995) to control the false discovery rate (FDR) can be applied. This method is the least stringent of all the four methods, and therefore leads to less false negatives at the cost of more false positives. It is defined by the following: Let S be the total number of features called significant, a sensible balance between the number of false positives F and the number of correctly rejected null hypotheses T can be achieved by considering

$\frac{number\, of\, false\, positive\, features}{number\, of\, significant\, features} = \frac{F}{F+T} = \frac{F}{S}$.

The FDR is defined to be the expected value of this quantity $FDR = E(\frac{F}{S})$. To control for multiple testing, let $p_1 \leq ... \leq p_k \leq ... \leq p_{n_{feat}}$be the ranked p-values with respective null hypotheses $H_{0,1}, ..., H_{0,k}, ..., H_{0,n_{feat}}$. As long as $p_k \leq \frac{k}{n_{feat}}\alpha$ the null hypotheses $H_{0,1}$ to $H_{0,k}$ can be rejected.

For completion, various algorithms exist for the analysis of differential expression incorporating multiple testing correction. Among the most widely used methods are the significance analysis of microarrays (SAM) which estimates the false discovery rate (*Tusher et al.*, 2001), linear models for microarray analysis (LIMMA) which uses adjusted p-values to control the FWER (*Smyth*, 2004), empirical Bayes analysis of microarrays (EBAM) which uses a simple nonparametric empirical Bayes model and where the empirical Bayes inferences are closely related to the FDR criterion (*Efron and Tibshirani*, 2002) and the method of rank products (RankProd) which provides a non-parametric approach to determine the significance level for each gene and which allows the control of the FDR and FWER (*Breitling et al.*, 2004).

2.4.3. Classification methods

In general, there are two different classification methods which can be distinguished, unsupervised and supervised classification. *Unsupervised classification* disregards prior knowledge and can be applied for the identification of sample subgroups such that objects within each subgroup are more closely related to one another than objects assigned to different subgroups (*Quackenbush*, 2006a). Different methods include hierarchical cluster analysis and principal component analysis (PCA). As the focus of this thesis is on supervised classification, further information on unsupervised classification can be found elsewhere (*Quackenbush*, 2006b; *Allison et al.*, 2006). In *supervised classification*, also known as class prediction, sample classes in a training set are known *a priori* and a reusable predictive model can be build from this training set to classify future samples into the predefined output classes. Many algorithms have been applied to high-throughput expression data. Among the most prominent methods are k-Nearest neighbor classifier, decision trees, neural networks, quadratic and linear discriminant analysis, shrunken nearest centroids, random forest and support vector machines. All formulas used in this section are taken from (*Tarca et al.*, 2007; *Hastie et al.*, 2001) if not stated otherwise.

2.4.3.1. Terminology

In this section, vector and matrix notation (x denotes an ordered p-tuple of numbers for some integer p, X denotes a rectangular array of numbers with x_{ij} being the value in the i-th row and j-th column of X) is employed.

In *supervised learning*, a set of *input* variables is used to predict the values of one or more *outputs*. In contrast to *regression* analysis where the output is *quantitative*, *qualitative* outputs are predicted in classification analysis.

Generally, a collection of input variables $i = 1, ..., n$ is classified into K predefined classes derived from a finite set. For a binary classification task, K is derived from $[0, 1]$ representing the two distinct classes case and control. An available set of measurements can then be organized in a $n \times p$ matrix $X = (x_{ij})$, where x_{ij} represents the observed values of the variable (e.g. feature intensity) j in the independent object (e.g. RNA sample) i. To every row of the matrix X, which is a vector x_i with p features, a response (class label) $y_i, y = 1, ..., c, ..., K$ is associated. A prediction rule (model) $C(x)$ based on the information from the input samples can be constructed. This classifier can be viewed as a collection of K discriminant functions $g_c(x)$ that partition the feature space X into K disjoint subsets $A_1, ..., A_K$ such that for an object x_i with measurements $x_i = (x_{i1}, ..., x_{ip}) \epsilon A_K$ the predicted class is k. Classifiers are build from a known learning or training set $\mathcal{L} = \{(x_1, y_1), ..., (x_n, y_n)\}$ and may be applied to a test set $\mathcal{T} = \{(x_1, y_1), ..., (x_{nT}, y_{nT})\}$ where each future unlabelled observation x_0 can be predicted by this rule $C(x_0|X) = y_0$.

2.4.3.2. Quadratic and linear discriminant analysis

This standard classification approach assumes that for each class c, x follows a multivariate normal distribution $N(m_c, \Sigma_c)$ with mean m_c and covariance matrix $\Sigma_c$where the element i, j of this matrix is the covariance between the variables i and j. Then the discriminant function for each class can be written as

$$g_c(x) = -(x - m_c)\hat{\sum}_c^{-1}(x - m_c)^T - log(|\hat{\sum}_c|) \qquad (2.4.4)$$

with $m_c = \frac{1}{n_c}\sum_{i=1}^{n_c} x_i$ and $\hat{\Sigma}_c = \frac{1}{n_c}\sum_{i=1}^{n_c}(x_i - m_c)^T(x_i - m_c)$.

The discriminant functions yield higher values for larger densities $p(x \mid y = c)$ and differences are only based on estimates of the mean and covariance matrix. The class of a new object x_0 will be defined on its largest discriminant value. As class boundaries here are nonlinear (quadratic), this method is called **quadratic discrimant rule**.

Alternatively, **linear discriminant analysis** (LDA) is based on the assumption that all the classes have the same covariance matrices $\Sigma_c = \Sigma \; \forall k$. Then a single pooled covariance matrix is used. Another assumption is that the data must be linear separable which suggests that the groups can be separated by a linear combination of features that describe the objects

These classifiers work optimally when their underlying assumptions are fulfilled which is not the case in many applications.

2.4.3.3. k-Nearest neighbor classification

The k-NN classification is not based on prior data distribution assumptions and requires no model fitting to the data which makes it an easy method. For each new object x_0, the k nearest vectors from the training set $x_i, i = 1, ..., n$ are found based on distance measures. A popular distance metrix is the Euclidean distance

$$d_{euc}(x, z)) = \sqrt{\sum_{j=1}^{p}(x_j - z_j)^2} \tag{2.4.5}$$

Then the classification is done by majority voting among the k neighbors and put simply, x_0 is assigned to the class most highly represented in its k nearest neighbors. When n_c denotes the number of objects among the k ones which belong to the class c, then the k-NN method classifies x_0 in the class that maximizes n_c, hence $g_c(x) = n_c$.

2.4.3.4. Nearest shrunken centroid classification

Nearest shrunken centroid (NSC) classification, also known as prediction analysis of microarrays (PAM), is an enhancement of the nearest centroid classification method (*Tibshirani et al.*, 2002). Briefly, the square distance from a test sample x_0 to each of the class centroids is computed and the class whose centroid is the closest is the predicted class. NSC classification modifies the standard nearest centroid classification. The idea is to shrink each class centroid toward the overall centroid for all classes by an fixed amount (Fig. 2.4.2) .

In details, the method works as follows:

Let x_{ij} be the measurement for feature $i = 1, ..., p$ and sample $j = 1, ..., n$ and C_k be the indices of the n_k samples in class k. Then the i-th component of the centroid for class k is defined as the mean expression in class k for feature i :

$$\bar{x}_{ik} = \sum_{j \in C_k} \frac{x_{ij}}{n_{k\ ij}} \tag{2.4.6}$$

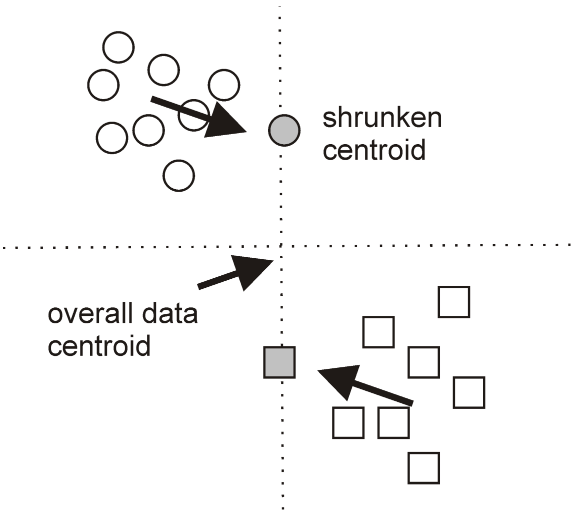


Figure 2.4.2.: Nearest shrunken centroids classification
NSC represents each class by its centroid and classifies new instances by assigning them the class of the closest centroid. NSC shrinks the class centroids in the direction of the overall data centroid. The figure was modified from *Struyf et al.* (2008).

and the i-th component of the overall centroid is

$$\bar{x}_i = \sum_{j=1}^{n} \frac{x_{ij}}{n} \tag{2.4.7}$$

The class centroids are shrunken toward the overall centroids after standardizing by the within-class standard deviation for each feature. This normalization is defined by a t statistic comparing class k to the overall centroid

$$d_{ik} = \frac{(\bar{x}_{ik} - \bar{x}_i)}{s_i} \tag{2.4.8}$$

where s_i is the pooled within-class standard deviation for feature i :

$$s_i^2 = \frac{1}{n-K} \sum_{k} \sum_{i \in C_k} (x_{ij} - \bar{x}_{ik})^2 \tag{2.4.9}$$

Then each d_{ik} is shrunken towards zero, giving d'_{ik}. Here, soft thresholding is used as shrinkage which is defined by

$$d'_{ik} = sign(d_{ik})(\mid d_{ik} \mid - \Delta)_{+} \tag{2.4.10}$$

where + means the positive part ($t_{+} = t$ if $t > 0$ and zero otherwise) meaning each d_{ik} is reduced by an amount Δ in absolute value, and is set to zero if its absolute value is less than zero. The parameter Δ is selected by cross-validation which is described in section sec. 2.4.4. If the shrinkage parameter Δ is large enough, many features are eliminated as far as class prediction is concerned. By standardization, higher weights are given to features with stable patterns within the same class.

This procedure yields the new shrunken centroids or prototypes

$$\bar{x}'_{ik} = \bar{x}_i + s_i d'_{ik} \tag{2.4.11}$$

After shrinking the centroids, the new sample x^* is classified by the usual nearest centroid rule, but using the shrunken class centroids. This classification rule is

$$C(x^*) = l \text{ if } \delta_l(x^*) = min_k \delta_k(x^*) \tag{2.4.12}$$

where δ_k are the discriminant scores, similar to those used in LDA, for class k

$$\delta_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \overline{x}'_{ik})^2}{s_i^2} - 2log\pi_k \tag{2.4.13}$$

2.4.3.5. Support vector machines

A support vector machine (SVM) discriminates one class form the other by a separating hyperplane with maximum margin (*Furey et al.*, 2000; *Vapnik*, 1998). Since there are many possibilities for a decision boundary that is capable of separating a binary linearly classification problem correctly (Fig. 2.4.3A), SVMs find the one that achieves the greatest margin between the two classes (Fig. 2.4.3B), especially to objects close to the margin. Those are called support vectors.

Let $\mathcal{L} = (x_1, y_1), ..., (x_{N_T}, y_{N_T})$ be the labeled training set with $x_i \in \Re^p, y_i \in \{-1; +1\}$. The hyperplane is defined by

$$\{x : f(x) = wx^T + b = 0\} \tag{2.4.14}$$

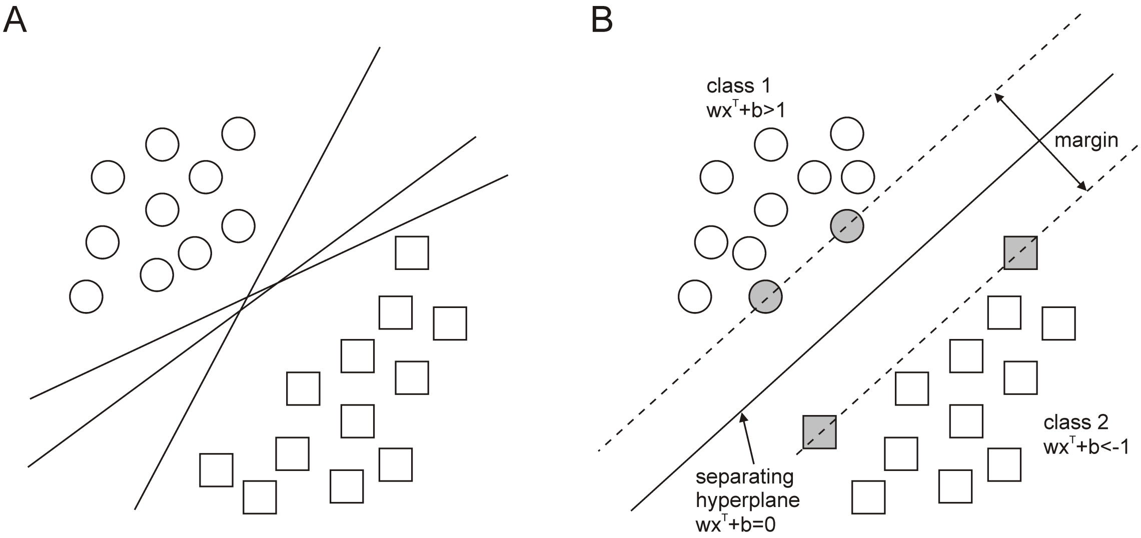


Figure 2.4.3.: Support vector machines class boundaries
A: A binary classification problem (two classes marked with circles and squares) with different linear decision boundaries. B: The maximum-margin decision boundary implemented by the SVMs together with the separating hyperplane. Samples along the dashed lines marked in grey are called SVs. The figure was adapted from *Tarca et al.* (2007).

where w is the p-dimensional vector perpendicular to the hyperplane with $\|w\| = 1$ and b is the bias. SVMs find the optimal w and b such that the hyperplane separates the data and maximizes the margin $1/\|w\|^2$ (Fig. 2.4.3B). The linear SVM problem can be formulated by introducing non-negative slack variables ξ_i, a penalty function measuring classification errors and a penalty parameter C set by the user as follows:

$$min_w(\frac{1}{2}\|x\|^2 + C\sum_{i=1}^{N_T}\xi_i) \tag{2.4.15}$$

subject to constraints

$$y_i(wx_i^T + b) - 1 + \xi_i \geq 0, \forall i \tag{2.4.16}$$

The classification function can be defined by

$$f(x) = sign(wx^T + b) = sign(\sum_i \alpha_i y_i (x_i x^T) + b). \tag{2.4.17}$$

The solution to this dual problem and the computation of the coefficients α_i can be found using quadratic programming.

It is not always possible for a single linear function to completely separate two given sets of points (Fig. 2.4.4A). In case where more sophisticated decision boundaries are needed, SVMs use the so-called *kernel trick* to still fit a hyperplane to the data. Therefore, the data points are projected by the use of kernel transformations into a higher-dimensional feature space where the data points effectively become linearly separable in the transformed space (Fig. 2.4.4B). The decision boundaries are linear in the projected feature space whereas the constructed hyperplane is nonlinear in the original input space (Fig. 2.4.4C).

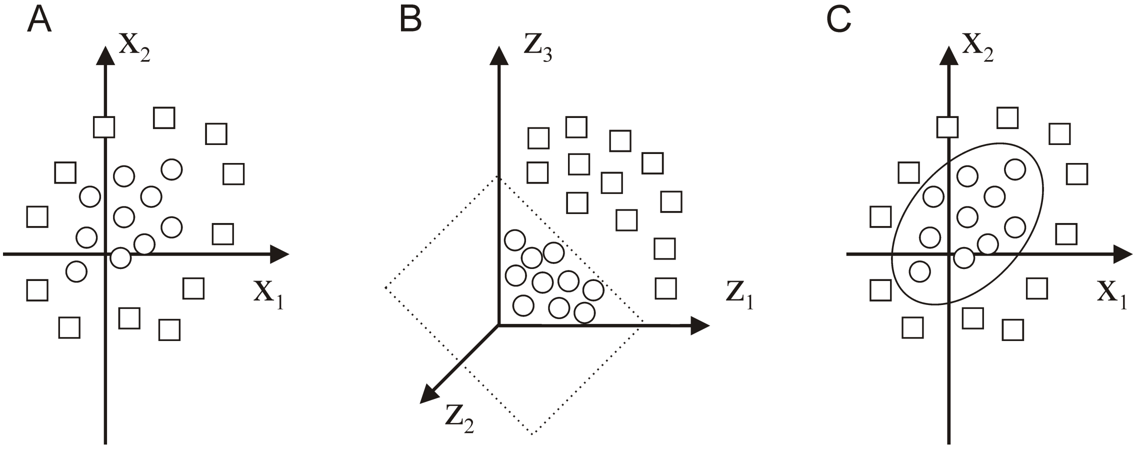


Figure 2.4.4.: The support vector machine kernel trick
When objects are not linearly separable (A), the SVM approach uses the kernel trick to transform the vector space into a higher dimensional feature space in which a linear hyperplane can be fit to the data points (B). The construced hyperplane could be nonlinear in the original input space (C). The figure was adapted from *Van Looy et al.* (2007).

The kernel transformation is performed by replacing every matrix product $\left(x_i x^T\right)$ with a nonlinear kernel function $K\left(x_i x\right)$. Common choices for K are

- linear: $K(x,z) = x^T z$
- polynomial: $K(x,z) = (x^T z + 1)^d$
- radial: $K(x,z) = e^{(-\gamma\|x-z\|^2)}$
- sigmoidal: $K(x,z) = tanh\left(\gamma x^T z + c_0\right)$

where d, γ and c_0 are parameters to be tuned to get better performance, e.g. using cross-validation.

2.4.4. Resampling designs

The best approach for model assessment would be to divide the data set into three parts: a training, validation and test set. The training set is used to build the models; the validation set is used to estimate the model performance and to choose the best one; the test set should only be used to assess the performance of the final chosen model. Independent validation means that there is no overlap between the training and testing data sets. In the absence of a large, independent and blinded test set (**external validation**), which is often due to the paucity of microarray data, numerous techniques exist for assessing the prediction accuracy by implementing some form of partitioning or resampling of the original observed data (**internal validation**). Each of these techniques involves dividing the data into a learning or training set $\mathcal{L}$ and a test set $\mathcal{T}$ to assess the predictive ability of the trained model. If the same data is used in both $\mathcal{L}$ and $\mathcal{T}$, referred to as *resubstitution*, the performance will be optimistically biased (*Hastie et al.*, 2001). Sample splitting, cross-validation and the bootstrap are fundamental tools for efficient sample re-use and work as follows:

- ***k*-fold Cross-validation** (CV): This method randomly systematically splits the data into k approximately equal-size subsets. Each partition is used as a test set for the model build on its complement, the remaining $k-1$ subsets. The average of the k prediction estimates forms the k-fold CV estimates (*Golub et al.*, 1979; *Speed*, 2003). A common choice for k is 10 (10-fold CV). An extreme form of CV is the so called leave-one-out CV (LOO-CV). Here, a data set of size n leads to n model fits, each using $n-1$ records ($k = n$). LOO-CV is almost unbiased at the cost of high variance.
- **Sample splitting**: Data splitting also known as the holdout method (*McLachlan*, 1992) entails a single division of the data into a training and test set based on a predetermined p using the test set to assess predictive ability of the trained model. For example, $p = \frac{1}{3}$ would lead to a ratio 2:1 for training and test set.
- **Bootstrap**: The bootstrap is a general tool for assessing statistical accuracy introduced by (*Efron and Tibshirani*, 1993). Basically, data sets are randomly drawn with replacement from the training data, each bootstrap sample the same size as the original training set. This is done B times, producing B bootstrap data sets. The model is fit to each of the bootstrap sets and examined over the B replications. Due to the overlap between training and test set, predictions can be overestimated. The ".632+ estimator" can be used to take into account the amount of overfitting (*Efron and Tibshirani*, 1997).

Using multiple resampling, e.g. with 100 repetitions of each method, one can obtain a mean, as well as a standard deviation, for the classifier performance.

2.4.5. Performance measurements

A number of different measures are commonly used to evaluate the performance of predictive algorithms. These measures differ according to whether the output is derived from a predefined set of responses or the classifier result is a quantitative score. A binary classification task is to distinguish between positive and negative samples and its performance can be assessed by the following metrics (*Shapiro*, 1999).

When the diagnostic test results in classification into positive and negative samples, the counts of correct and incorrect predictions can be summarized in a 2 × 2 contingency table or confusion matrix (see Tab. 2.3) of predictions against actual class labels. Most commonly, the terms sensitivity and specificity are used to characterize a classification rule and can be calculated from the entries of this contingency table.

	Predicted positive	Predicted negative	
Actual positive	TP	FP	PPV
Acutal negative	FN	TN	NPV
	PP	PN	

Table 2.3.: Contingency table or confusion matrix
TP: true positives (predicted positive, actual positive); TN: true negatives (predicted negative, actual negative); FP: false positives (predicted positive, actual negative); FN: false negatives (predicted negative, actual positive); PPV: positive predictive value; NPV: negative predictive value; PP: predicted positives (sensitivity); PN: predicted negatives (specificity).

Sensitivity (sens) or true positives (TP) measures the probability of predicting positives given true positive status, for example the probability that a positive sample is predicted to be truly positive. It is calculated by

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.4.18}$$

Specificity (spec) or true negatives (TN) relates the test's ability to identify negative results as the proportion of controls that will test negative for it. It can be written as

$$Specificity = \frac{TN}{TN + FP} \tag{2.4.19}$$

The total **accuracay** (proportion of correct prediction) from the contingency table can be calculated by

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.4.20}$$

As accuracy is sensitive to the prior class probabilities and does not fully describe the actual difficulty of the decision problem for highly unbalanced distributions, the **Matthews correlation coefficient** (MCC) can be used as a balanced performance measurement. It can be interpreted as a correlation coefficient between observed and predicted binary classifications and as for the Pearson correlation, a value of 1 corresponds to a perfect correlation, meaning a perfect performance. The MCC is defined as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.4.21}$$

Some classification models result in a continuous output (e.g., an estimate of an instance's class membership probability or another quantitative score) that represents the degree to which an object is a member of the specific class. When assessing the distribution of test results X for positive and negative groups, the degree of overlap determines a tests discriminatory ability and by introducing a threshold or decision limit c, the samples can be separated into predicted negatives ($X > c$) and predicted positives ($X > c$). (see Fig. 2.4.5).

Different discrimination thresholds of c can then be applied to predict membership into the two classes and yields a different 2×2 contingency table where true positive rate (TPR) and false positive rate (FPR) can be estimated. When c increases, sensitivity increases as well at the cost of reduced specificity and vice versa, a decreasing c leads to higher specificity with decreasing sensitivity.

The **receiver operating characteristic** (ROC) curve is a useful graphical plot to visualize classifier performance for such a varying threshold c (*Swets*, 1988). The ROC curve is constructed by using different values of the threshold c to plot the Sensitivity ($sens(c)$) on the y-axis against 1-specificity ($1 - spec(c)$) on the x-axis (see Fig. 2.4.6). The point $(0, 1)$ represents perfect classification with 100% sensitivity and 100% specificity whereas a random model guessing class labels would lead to a ROC curve at the diagonal line.

The information in the ROC curve can be reduced to one single scalar summary metric of predictive performance, the **area under the ROC curve** (AUC) (*Bradley*, 1997). An AUC value close to 1 indicates excellent classification whereas a value of 0.5 indicates useless prediction performance. The AUC is a robust measure of performance and compared to the MCC, it is independent of the choice of the threshold c.

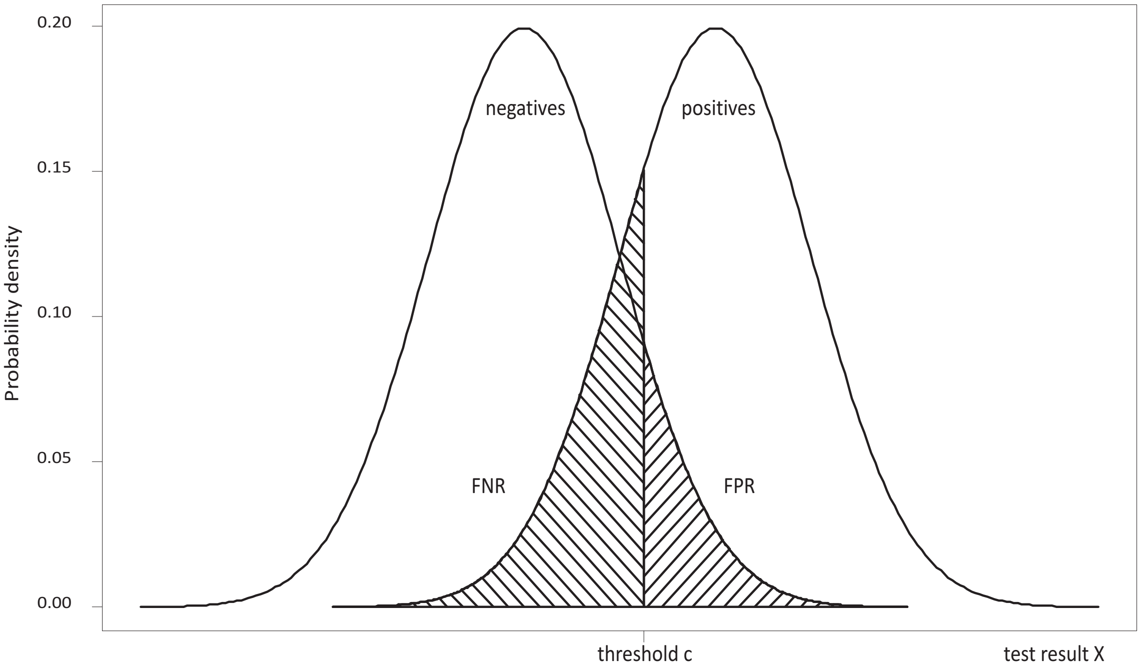


Figure 2.4.5.: Hypothetical distributions of diagnostic test results X for negative and positive samples. The vertical line at the threshold $X = c$ indicates the decision limit for a positive test. The shaded area to the right of c is the false positive rate (FPR); the shaded area to the left of c is the false negative rate (FNR). The figure was modified from *Shapiro* (1999).

The AUC is equal to the value of the Wilcoxon-Mann-Whitney test statistic and also the probability that the classifier will rank a randomly drawn positive sample higher than a randomly drawn negative sample ($AUC = Prob(positive > negative)$). Furthermore, the AUC represents the average sensitivity over all values of FPR.

ROC curves and the AUC can be estimated under parametric or non-parametric assumptions as described in (*Faraggi and Reiser*, 2002; *Shapiro*, 1999). The non-parametric approaches include the use of the Mann–Whitney statistic and the fit a smooth ROC curve using kernel smoothing followed by estimation of the AUC by integration. The parametric approaches cover the assumption that the marker values for negative and positive samples are normally distributed where the AUC can be estimated by parametric methods as well as the application of a Box–Cox type power transformation together with the use of normal theory.

Additional to the AUC, the **Youden Index** (*Youden*, 1950) is frequently used in practice. This index is defined as

$$J = max_c \{sens(c) + spec(c) - 1\} \tag{2.4.22}$$

and ranges between 0 and 1. The Youden Index (YI) has an attractive feature

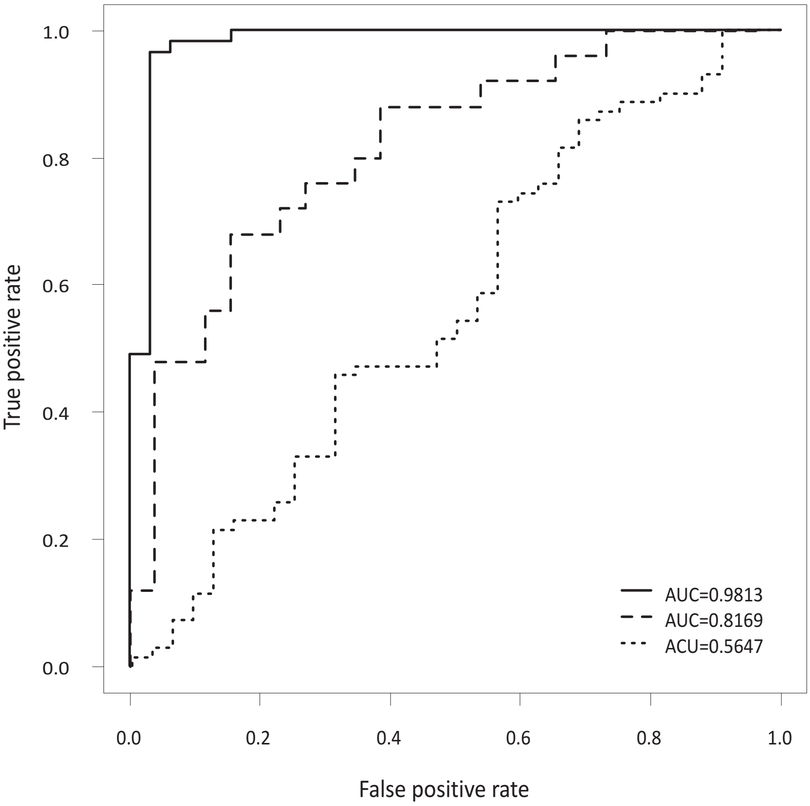


Figure 2.4.6.: Examples of ROC curves
Shown is the true positive rate on the y-axis against the false positive rate on the x-axis for three different ROC curve examples (AUC=0.9813 excellent, AUC=0.8169 good and AUC=0.5647 worthless classification)

not present in the AUC: it enables the choice of an optimal threshold value c^*, the threshold value for which $sens(c) + spec(c) - 1$ is maximized. The YI is easy to apply and does not require further information such as prevalence rates.

2.5. Epidemiology and clinical characteristics of analyzed diseases

2.5.1. Cancer

Cancer is a generic term for a class of diseases characterized by unregulated cell growth. Abnormal cells grow beyond their usual boundaries and can invade adjoining parts of the body. Uncontrolled dividing cells form a lump or masses of tissue called tumor. Not all tumors are concerous, benign tumors can be removed and do not spread to other parts of the body. In contrast to this, malignant tumors spread throughout the blood or lymph system to to other organs, a process known as metastasis (*National Cancer Institute*, 2012; *American Cancer Society*, 2012; *Cancer Research UK*, 2012). There are over 100 different types of cancer and each subtype is classified by the type of cell or organ that resembles the tumor (*National Cancer Institute*, 2012). Broader categories of cancer include:

- Carcinoma: Malignant tumors derived from skin or in tissues that line or cover internal organs. This group represents the most common type of cancer, including the common forms of breast, prostate, lung and colon cancer.
- Sarcoma: This rare cancer type arises from bone, cartilage, fat, muscle, blood vessels, or other connective or supportive tissues.
- Leukemia: Type of cancer of the blood or bone marrow. It is characterized by the production of large numbers of abnormal blood cells that enter the blood and do not form tumors.
- Lymphoma and myeloma: Malignancies derived from the cells of the immune system.
- Central nervous system cancer: Type of cancer that begin in the tissues of the brain and central spinal cord.

Cancer is a leading cause of death worldwide. In 2008, the World Health Organization (WHO) estimated about 12.7 million cancer cases and 7.6 million cancer deaths (13% of all deaths) to have occurred (*Jemal et al.*, 2011; *World Health Organization*, 2012).

2.5.1.1. Lung cancer

Lung cancer is a disease of uncontrolled cell growth that forms in tissues of the lung. This growth may lead to metastasis and impede the function of the lung. Symptoms of lung cancer include shortness of breath, unexplained weight loss, chronic fatigue, and coughing (*National Cancer Institute*, 2012).

Lung cancer is still the leading cause of cancer related death worldwide. It is the most common cause of cancer-related death in men and the second most common in woman with a total of 1.37 million deaths (18% of all deaths) in 2008. Tabacco use is the most important risk factor for lung cancer causing 71% of global lung cancer deaths, other risk factors include genetic factors, asbestos, radon, and air pollution (*Jemal et al.*, 2011; *World Health Organization*, 2012).

The two main types are small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), both categories refer to the size and appearance of the malignant cells seen by a histopathologist under a microscope. SCLC accounts for about 13% of all lung cancers and is an aggressive and fast-growing cancer that spreads quickly. The cancer cells look small and oval-shaped when looked at under a microscope. NSCLC is the most common kind of lung cancer (about 87%) and spreads more slowly than SCLC (*National Cancer Institute*, 2012). Lung cancers are described in different stages, starting from an occult stage 0 where the cancer has not invaded nearby tissues to stage IV where malignant growths of cells may be found in more than one area of the other lung, the fluid surrounding the lung, or distant parts of the body.

Usually, diagnosis is performed by chest radiography and computed tomography (CT) imaging to gain more information about the type and extend of the disease. In a biopsy, a pathologist examines lung tissue samples under a microscope to make a definite diagnosis (*National Cancer Institute*, 2012). The type of lung cancer and the stage of the disease determine what type of treatment is needed. Prognosis has remained poor with a disastrous 5-year survival rate of only about 15.9% due to diagnosis of the disease in late, i.e. incurable stages in the majority of patients (*Jemal et al.*, 2008; *National Cancer Institute*, 2012). The 5-year survival rate for cases detected with a early, localized stage of disease is 52%, however only 15% of lung cancers are diagnosed at this early stage. NSCLC has a higher 5-year relative survival rate (17%) in comparison to SCLC (6%) (*American Cancer Society*, 2012).

2.5.1.2. Multiple myeloma

Multiple myeloma (MM) is a cancer of plasma cells, a type of white blood cells, characterised by clonal proliferation of malignant plasma cells in the bone marrow microenvironment (*Palumbo and Anderson*, 2011). These abnormal plasma cells accumulate in the bone marrow and may damage the solid part of the bone (*National Cancer Institute*, 2012). MM has many varying symptoms including bone pain, frequent infections, and weight loss. MM is diagnosed with blood tests and bone morrow examination. Additional studies include standard cytogenetic analysis as any chromosomal abnormality is associated with a worse outcome in comparison to a normal karyotype (*Palumbo and Anderson*, 2011). MM characterizes a heterogeneous picture in terms of symptoms, biologic characteristics, response to treatment, and clinical outcome and several studies focus on translating the wide molecular heterogeneity into classification and prognosis models that could be used for patient management (*Zhan et al.*, 2006; *Avet-Loiseau et al.*, 2009; *Broyl et al.*, 2010; *Munshi and Avet-Loiseau*, 2011).

MM is the second most common hematological malignancy and accounts for approximately 1% of neoplastic diseases and 13% of hematologic cancers (*Palumbo and Anderson*, 2011) and most people are diagnosed after the age of 65 (*National Cancer Institute*, 2012). The 5-year relative survival rate for MM is around 40% (*American Cancer Society*, 2012) and survival is lower in elderly people. Patients under 60 years have a 30% 10-year relative survival rate (*Brenner et al.*, 2008).

2.5.1.3. Leukemia

Leukemia is a malignant disease of the myeloid white blood cell line characterized by a increased growth of abnormal white blood cells. Malignant cells interfere with normal blood cell production and initially accumulate in bone marrow and peripheral blood with further suppression of its normal counterparts and later infiltrate solid organs and tissues (*National Cancer Institute*, 2012). To understand the different types of leukemia, a basic overview of the blood and lymph system is given below:

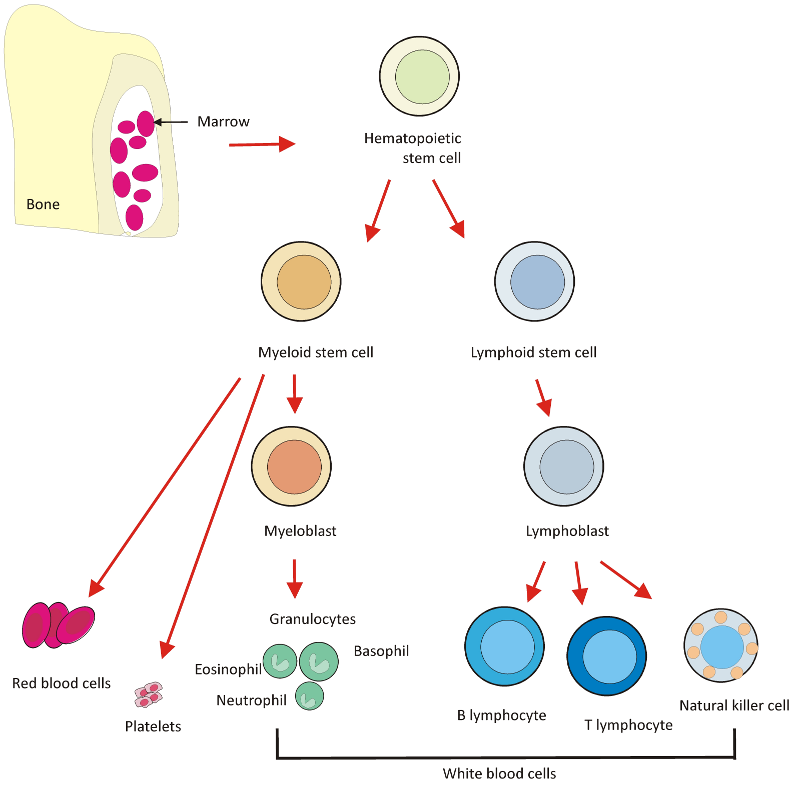


Figure 2.5.1.: All cells of the immune system arise from hematopoietic stem cells in the bone marrow
The pluripotent HSCs divide to produce several types of progenitor cells, the lymphoid stem cell and the myeloid stem cell. The lymphoid progenitor cell gives rise to lymphocytes, including T lymphocytes, B lymphocytes, and natural killer cells. The myeloid progenitor cell gives rise to, for example, erythrocytes (red blood cells), platelets, and granulocytes. The figure is adapted from (*National Cancer Institute*, 2012).

Most blood cells are developed from cells in the bone marrow called hematopoietic stem cells (HSCs). Pluripotent HSCs mature into different kinds of blood cell types from the myeloid and lympoid lineages (Fig. 2.5.1). Myeloid stem cells develop into red blood cells (erythrocytes), platelets and granulocytes (white blood cells). Red blood cells mainly carry oxygen to other tissues of the body, platelets help prevent bleeding and granulocytes fight infections and diseases. This system responds early and nonspecifically to infection. The lympoid stem cells develop into a lymphoblast cell and then differentiates into one of the three types of lymphocytes (white blood cells): B lymphocytes (B cells), T lymphocytes (T cells), and natural killer cells. Upon immune activation, B cells differentiate into plasma cells that produce and release thousands of specific antibodies into the bloodstream. The T cells differentiate into cells that can kill infected cells or activate other cells of the immune system, thereby coordinating the entire immune response. Natural killer cells attack cancer cells and viruses (*National Cancer Institute*, 2012). White blood cells, red blood cells, and platelets are produced on demand and older or damaged cells die.

Leukemia is subdivided into a chronic and an acute form. Chronic leukemia is characterized by slow cancer progression and in early stages of disease, abnormal blood cells still retain parts of their function and people may not have any cancer symptoms. Most of the abnormal cancer cells are partly mature but not completely and look more like normal white blood cells. In acute leukemia, the number of abnormal blood cells increases rapidly and the disease progresses quickly. Acute leukemia develops from immature blood cells that cannot carry out their normal function. Leukemia can arise in lympoid or myeloid cells and taken together, these two classifications provide four main types: Acute myeloid leukemia (AML), chronic myeloid leukemia (CML), acute lymphocytic leukemia (ALL), and chronic lymphocytic leukemia (CLL).

The production of normal platelets, red and white blood cells often is stopped from the bone marrow. Deficiency of white blood cells impairs a patient's ability to fight infections and therefore, recurrent minor infections or poor healing of minor cuts because of inadequate white blood cell counts may occur. A shortage of red blood cells results in a condition called anemia, whose symptoms include paleness and easy fatigue. The lack of platelets can lead to easy bleeding and bruising (*American Cancer Society*, 2011). People who have chronic leukemia may not have major symptoms; they may be diagnosed as a result of a routine physical examination. Leukemia strikes males and females of all ages and its cause of most leukemia cases is unknown (*American Cancer Society*, 2012), however, leukemia risk factors include exposure to high levels of radiation and exposure to high levels of benzene.

Leukemia accounts for 350,000 cancer cases worldwide in 2008 (*Cancer Research UK*, 2012). From estimated 274,930 leukemia cases in the United States in 2008, ALL accounted for approximately 58,854 cases, CLL for 105,119 cases, AML for 30,993 cases and CML for 26,359 cases (*American Cancer Society*, 2011). The most common leukemia types in adults are AML and CLL, whereas ALL accounted for 76 percent of new leukemia cases in children and adolescents ages 0 to 19 years. Furthermore, leukemia is the most common cancer in children and adolescents less than 20 years old. Approximately 21,780 people died from leukemia in the United States in 2011.

From 2001 to 2007, the overall leukemia 5-year relative survival rate in the United States was 57%. In detail, the 5-year survival rates were 55% for CML, 81% for CLL, 24% overall and 64% for children and adolescents younger than 15 years in

AML, and 67% overall and 91% for children and adolescents younger than 15 years in ALL (*American Cancer Society*, 2011).

The diagnosis of leukemia requires specific blood tests and bone marrow examination to determine the type of cells and the number of mature cells and blasts (*National Cancer Institute*, 2012). Importantly, current diagnostic strategies involve bone marrow examination by an experienced hematologist, flow cytometry, polymerase chain reaction (PCR) as well as cytogenetic analysis (*Cheson et al.*, 2003). This rather time-consuming procedure does not allow for designing novel induction therapies according to risk factors.

2.5.2. Human immunodeficiency virus

The Human Immunodeficiency Virus (HIV) is a retrovirus that infects cells of the immune system and therefore, destroys of impairs a person's surveillance and defense system against infections. Infected persons become immunodeficient, resulting in increased susceptibility to a wide range of infections and diseases that a healthy immune system normally can fight off. The most advanced stage of HIV infection is Acquired Immunodeficiency Syndrome (AIDS). It can take 10-15 years for an HIV-infected person to develop AIDS (*World Health Organization*, 2012).

HIV symptoms depend on the stage infection, starting with flu-like illnesses and then with progressing infection and a weaker immune system, symptoms may include weight loss, fever, and cough. Without treatment, severe illnesses such as tuberculosis, meningitis and cancers, can develop. HIV is transmitted through unprotected contact with a variety of body fluids of infected individuals. Diagnosis of HIV infection can be done by blood tests detecting presence or absence of antibodies and antigens in the blood (*World Health Organization*, 2012). HIV can be suppressed by combination antiretroviral therapy (ART). The HIV infection is not cured by ART, but the viral replication is controlled and people can live healthy with HIV.

The WHO reported that HIV caused more than 30 million deaths over the past three decades. In 2011, approximately 34.2 million people were living with HIV, over 60% of infected people live in Africa. A total of 2.7 million people were newly infected with HIV in 2010 and 1.8 million died from AIDS in 2010 (*World Health Organization*, 2012).

After HIV infection, the individual genetic background plays a central role for the variable disease progression towards AIDS. Long-term non progressors (LTNPs) are rare represented individuals in the HIV infected patient population ($1-2\%$) having AIDS-free survival without ART for more than 10-15 years (*Sheppard et al.*, 1993). It is estimated that 1 in 500 people with HIV maintain this favorable course of infection (*National Institute of Allergy and Infectious Diseases*, 2012). LTNPs are of interest because they serve as an ideal model for HIV vaccine development due to their natural control of HIV infection (*Poropatich and Sullivan*, 2011).

2.5.3. Tuberculosis

Tuberculosis (TB) is an infectious disease caused by *Mycobacterium tuberculosis* strains. TB bacteria usually attack the lungs. Most infections remain asymptomatic, termed latent TB, and about $5 - 10\%$ progress to active disease, which can be fetal if left untreated (*Lawn and Zumla*, 2011). Importantly, latent TB can turn into active TB at any time of the infected person's lifetime. People with compromised immune systems such as people with HIV have a higher risk of falling ill, 30% of TB patients coinfected with HIV develop active TB (*World Health Organization*, 2012).

Active TB may have mild symptoms such as cough, fever, and weight loss in the first months. Later on, symptoms of active TB are cough with sputum and blood, chest pains, weakness, fever and high sweats. TB is transmitted through the air by minute aerosol droplets (*McNerney et al.*, 2012).

TB diagnosis is made by finding *M. tuberculosis* bacteria in a clinical sample and accurate TB infection and disease status is difficult (*Barry et al.*, 2009). Given limited laboratory capacity, often light microscopic examination of sputum specimens is the only TB test available (*McNerney et al.*, 2012). This test performs poorly in young and immunocompromised people and fails to detect active disease cases. Hence optimal detection of active and latent TB infection, especially among HIV infected people, remains a major challenge in resource-limited settings (*McNerney et al.*, 2012).

One third of the world's population, approximately 2 billion people, is estimated to be infected with TB. The WHO indicated a total of 8.8 million incident cases of TB worldwide and 1.4 million associated deaths. After HIV/AIDS it is causes most deaths to a single infectious agent worldwide. About half a million children under 14 years have TB and approximately 64,999 children died from the disease in 2010 (*World Health Organization*, 2012). Most cases occur in Asia and Africa. About 13% of all TB cases occurred among people infected with HIV (*McNerney et al.*, 2012).

3. Material and Methods

3.1. Sample and data collection

For the analyses performed in this thesis, new data was generated in our laboratory and also previously published datasets were used. A detailed description of data sets is provided in the supplemental material (Appendix C). As to our knowledge all studies were performed following guidelines of the respective local Ethics Committees. For our own data, we followed the approval by the Ethics Committee of the University of Cologne.

3.1.1. Non-small cell lung cancer data set

NSCLC cases and hospital based controls were recruited at the University Hospital Cologne and the Lung Clinic Merheim, Cologne, Germany. Healthy blood donors were recruited at the Institute for Transfusion Medicine, University of Cologne. From all individuals blood samples stabilized using PAXgene^{TM}. Blood DNA tubes were taken for blood-based gene expression profiling. For all NSCLC cases blood was taken prior chemotherapy. In total, patients with NSCLC (n=95) of stage I-IV, hospital based controls (n=68) and healthy blood donors without prior history of lung cancer (n=70) were included. The hospital based controls enclosed individuals suffering from advanced chronic obstructive pulmonary disease (COPD) (n=13), hypertension (n=28) or other malignancies (n=27). Detailed information on cases and controls are summarized in Tab. 4.1. All gene expression profiling (GEX) samples (in total n=233) were generated using the Illumina Human WG6-V2 BeadChip microarray and were used to establish a whole blood derived GEX signature to detect patients with NSCLC. Data is provided in the NCBI GEO database (*Edgar et al.*, 2002) with the accession number GSE12771.

3.1.2. MAQC-II multiple myeloma data set

The multiple myeloma dataset provided by the MAQC-II study (n=554, GSE24080) was used for initial setup of the analytical approach described within this thesis. These samples were all derived from highly purified bone marrow plasma cells. Plasma cells were enriched by anti-CD138 immunomagnetic bead selection of mononuclear cell fractions of bone marrow aspirates in a central laboratory. All samples applied to the microarray contained more than 85% plasma cells. Dichotomized overall survival (OS) and event-free survival (EFS) were determined based on a two-year milestone cutoff. For these GEX samples the Affymetrix microarray Human Genome model HG-U133A microarray format was used.

3.1.3. Data set for the development of a primary diagnostic test for AML

GEX experiments performed on the Affymetrix microarray HG-U133A in the NCBI GEO database (*Edgar et al.*, 2002) and from other sources including our own unpublished data characterizing human disease conditions based on peripheral blood or bone marrow samples were collected and annotated. Our resulting data set incorporated both acute and chronic leukemia and other unrelated disease samples as well as healthy controls across 17 data sets comprising 2013 individual samples. Patients with AML (n=717), acute lymphoblastic leukemia (ALL, n=230), chronic myeloid and lymphoblastic leukemia (n=98), infectious diseases (n=262), coronary artery disease (n=101), Parkinsons disease (n=85), Colitis ulcerosa and Crohns disease (n=85), Huntingtons disease (n=17), post-infectious chronic fatigue syndrome (n=8), and 410 healthy controls from these studies were included (see Tab. 5.1 and Tab. A.1 for detailed information of studies).

A second AML data set was derived from GEX samples (n=2088, see Tab. 5.3) performed on the Affymetrix HG-U133 2.0 microarray collected from 15 individual studies and comprised patients with AML (n=1093), myelodysplastic syndrome (n=302), healthy controls (n=155) as well as other diseases (n=538). The unrelated disease cohort comprises patients with liver transplant (n=9), Alzheimer/ dementia (n=6), samples from the malaria vaccine trial (n=39), multiple sclerosis (n=240), juvenile idiopathic arthritis (n=180), Steven-Johnson syndrome (n=5) and peritonitis (n=59).

3.1.4. Data set established to detect patients with active tuberculosis

The tuberculosis (TB) data set comprised all samples (n=147) from a recent study using whole blood derived GEX signatures to detect patients with ATB (GSE19491). This study includes patients with ATB (n=54), latent TB (LTB, n=69) and control samples (n=24). Detailed information can be found in Tab. A.3. Patients with ATB were confirmed by laboratory isolation of *M. tuberculosis* on mycobacterial culture of a respiratory specimen. Latent TB patients were defined by a positive *M. tuberculosis* antigen-specific IFN-γ release assay (IGRA) result. In this study the Illumina platform utilizing Illumina Human HT12-V3 BeadChip arrays was applied.

3.1.5. Data set established to detect patients with HIV infection

We generated a data set of GEX samples (n=257), to identify a HIV infection cohort irrespective of viral load including both HIV+ and long-term non progressor (LTNP) patients in comparison to control samples (see Tab. A.2 for detailed information). Patients were recruited into this study at the HIV outpatient clinics at the University hospital Cologne. Patients were included based on virological and serological diagnosis to be HIV+ and have documented HIV infection since $\geq$1 year. LTNP patients (n=27) were defined as 1st never treated with anti-retroviral drugs, 2nd CD4$^+$ T-cells $\geq 500/\mu L$, 3rd HIV-RNA in plasma $\leq$ 2000 *copies*/*mL* and 4th CDC stage A. Patients progressing to AIDS (n=79) were defined as 1st documented CD4$^+$ T-cell decline $\leq 300/\mu L$ and 2nd CDC stage C. Control samples from HIV- donors were collected in our local blood bank. Healthy controls (n=88) and samples of patients with the other inflammatory diseases sepsis (n=34) and scleroderma (n=29) were included. From all individuals PAXgene stabilized blood samples were taken for blood-based GEX. All samples were processed using the Illumina Human HT12-V2 BeadChips.

3.2. Statistical and bioinformatic analysis

All statistical and bioinformatical analysis were performed using R software (*Ihaka and Gentlemen*, 1996) and packages from the bioconductor project (*Gentleman et al.*, 2004). Following recommendations made by the MAQC-II consortium, all methods are provided in accompanying R scripts (Appendix D).

3.2.1. Data preparatory steps

Samples were subjected to an extended quality check prior use and all GEX data presented in this thesis were of high quality. First, a visual inspection of pairwise array's spatial distribution of feature intensities was performed using scatter plots from all arrays of a data set. Overall, the correlation was required to be above 0.7. Next, the present call rate had to reach a threshold determined above 0.2. Third, homogeneity between the arrays and the overall sample distribution was visually analyzed by box and density plots.

When Affymetrix microarrays were used, all samples of a data set provided by cel-files were normalized using the MAS5 method, a method summarizing background correction, signal intensity and scaling calculations, from the affy package. For Illumina microarrays, expression values were quantile normalized using the limma package as advised by *Dunning et al.* (2008). Illumina microarray data was provided in text-files comprising transcript intensity together with a corresponding detection p-value. Transcripts with mean signal intensity below a calculated background signal were removed from the final data set. The background signal intensity was calculated by linear regression modelling between normalized expression values and corresponding detection p-values. The estimated signal intensity at a detection p-value of 0.05 was used as the background threshold. Background signal intensities were 6.8093 for TB, 65.8358 for HIV and 57.9289 for NSCLC data, respectively. In total, the Affymetrix data sets contain 22,283 transcripts, whereas the TB data profiled 19,080, the HIV data 13,386 and the NSCLC data 13,264 transcripts.

Due to our overall strategy including simulation and adaptive learning methods and to be able to better mimic expected clinical routine of subsequent data generation, which is naturally prone to batch-effects, we voted against batch-effect removal when collecting samples from one platform. When comparing HG-U133A and HG-U133 2.0 data, we filtered transcripts present on both arrays and normalized those together using quantile normalization.

3.2.2. Standard classification

Generally, a mathematical model for classifier development has two steps. First, genes that are differentially expressed (DE) between the experimental groups are selected and second, a classifier is constructed using those identified DE genes. As the preferred feature selection approach, we used a combination of FC and p-values based on t-tests (2-sided, unequal variance = Welch test, unpaired) between the two experimental groups. The p-values were adjusted for multiple testing using Benjamini-Hochberg correction as implemented in the R stats and multtest packages. The gene filtering was performed in the training set (TS) of samples to avoid overfitting. Usually, a FC of 2 and a p-value of 0.05 were used for DE calculation. For supervised classification the packages e1071 (SVM), pamr (PAM) and class (LDA) were used. We utilized a linear SVM kernel if not stated otherwise and default, untuned parameters. For performance assessment, AUC, MCC, sens and spec were used. AUC values were calculated using prediction probabilities as implemented in the ROCR package (*Sing et al.*, 2005) applying the non-parametric AUC estimation. MCC, sens and spec were calculated at the maximum Youden Index.

For standard classification, the classifier was built and optimized based on the TS using 10 times repeated 10-fold CV. In the internal CV, feature selection cut-offs and classification algorithms are determined according to the maximum mean AUC values reached. A classifier was then build using the respective cut-off and selected algorithm in the TS and then validated in the independent validation set (VS).

3.2.3. Random sample-split design

A sample-split resampling design was used for clinical trial simulation (TSA, see sec. 5.2.2). We employ a multiple random validation setup, where the data are repeatedly randomly divided into a TS and two independent VSs without replacement. We perform $n_{ss} = 10,000$ and $n_{ss} = 1000$ random samplings, each with a ratio of

1:1:1 for the size of training to the two validation sample sizes if not stated otherwise. Class proportions remained balanced. The classifier was built in the TS based on DE genes selected by statistical testing between the two experimental groups in the TS. This model was then applied to the two independent VSs and prediction performance was assessed by calculating AUC, MCC, sens and spec respectively. For completion, prediction performance was compared between our sample split design (without replacement) to classical bootstrap (with replacement).

3.2.4. Resampling feature selection

The resampling design described above can furthermore be used to select informative features and to train a good classifier. All different B subsamplings of the original data lead to B different signatures obtained by feature selection from the TS in the TSA. A consensus feature signature can be build from the various signatures by different aggregation methods of ranking the features from most important to least important. Each feature will obtain a rank and the mean rank can be computed as

$$\bar{r}_j = \frac{1}{B} \sum_{b=1}^{B} r_{j,b}. \tag{3.2.1}$$

This means, assign a rank $r_{j,b}$ to feature j in each sample split b. Then we can compute the mean rank of each feature j and use only the best features or the best ranking features. Most important features can then be used to train the classifier on the whole data set. Different ways of obtaining the rank are as follows:

1. Each feature can be ranked according to the number of times it is identified as differentially expressed in each TSA. By using subsets of the original data set, stable markers would be expected to appear more often than uninformative features in the signatures.

$$r_{j,b} = \begin{cases} 1 & \text{if the feature } j \text{ is DE in } b \\ 0 & \text{otherwise} \end{cases} \tag{3.2.2}$$

2. The procedure in (1) would lead to a stable biomarker signature, yet still stability alone is not a good quality measure as it needs to be assessed together with classification performance. In linear SVM classification, the absolute values of the weights of each feature $w_{j,b}$ can be regarded as the contribution of each feature:

$$r_{j,b} = \begin{cases} w_{j,b} & \text{if } j \text{ DE} \\ 0 & \text{otherwise} \end{cases} \tag{3.2.3}$$

3. All features can be ranked according to both the calculated p-value and FC in the subset, hence $r_{j,b} = pV_{j,b}$ or $r_{j,b} = FC_{j,b}$.

3.2.5. Negative controls

Two different negative control procedures were implemented as well. First, class labels were randomly assigned to each single data set. TSA (1000 iterations) was repeated using this non-predictable data set.

Second to test the final classifier specificity, the whole analysis of classifier building and application was repeated 1000 times by using random feature sets of equal size.

3.2.6. Heuristic approach for artificial sample generation

In biomarker studies, the number of features is typically larger than the sample size due to the paucity of high-throughput genomic data. The generation of more samples often is not possible or expensive and time-consuming. With this practical limitation in mind, an alternative might be the inclusion of artificial samples. The idea is to simulate new samples according to the distribution of given samples. Here, metric relationships from two randomly chosen samples were used to calculate new samples in a rather simple feature-by-feature approach by the following heuristics:

Randomly draw two samples y, z from the same experimental group. When $y < z$, generate ten new sample values for each feature by calculating

$$\begin{array}{ll}
(y+z)/2 & y-(z-y) \\
y+\frac{1}{2}(z-y) & z-\frac{1}{2}(z-y) \\
y-\frac{1}{10}(z-y) & z+\frac{1}{10}(z-y) \\
y-\frac{1}{5}(z-y) & z+\frac{1}{5}(z-y) \\
y-\frac{1}{2}(z-y) & z+\frac{1}{2}(z-y)
\end{array}$$

Those parameters were chosen to increase the variance within the data set and to decrease homogeneity in the simulated dataset.

3.2.7. Adaptive learning approach

Central to the adaptive learning approach is the integration of each new sample or case into the classifier to stepwise optimize the diagnostic assessment. In contrast to static initial best biomarkers, continuous incremental improvement is evaluated as follows (see Figure Fig. 5.2.13A):

The complete AML data set (n=2013) is divided into a set for adaptive learning (ALS) and two external validation sets (ExVS1, ExVS2), each of size 671 samples. In the adaptive learning process, an initial TS of size 50 is drawn from the ALS.

Internal classifier validation is performed by TSA (100 iterations), summarized in a mean AUC performance assessment and then a classifier based on the entire TS is applied to ExVS1 and ExVS2. Next stepwise, additional sample sets of size 20 are subsequentially drawn from the ALS and for each new sample set added to the initial TS, internal and external validation is performed based on the enlarged set (of size 70, 90, 110 etc.). To ensure adaptive learning is not depending on sample draw ordering, this procedure is repeated 100 times, leading to summarized mean of mean of 100x100 AUCs.

The adaptive learning approach is furthermore developed to determine minimal sample size required for validation cohorts based on a small pilot trial data. Based on the summarized mean of mean AUCs, a cubic smooth spline curve is fitted to these values (*Hastie and Tibshirani*, 1990). This spline function requires the AUCs x_i stated as $x_0 < x_1 < ... < x_n$ as the predictor value and the sample sizes y_i as the responses, $i = 1, ..., n$. The relation $Y_i = f(x_i)$ models these observations. The smoothing spline estimate $\hat{f}$ of the function f is defined to minimize

$$\sum_{i=1}^{n}(Y_i - \hat{f}(x_i))^2 + \lambda \int_{x_1}^{x_n} \hat{f}\text{"}(x)^2 dx \tag{3.2.4}$$

where $\hat{f}$ is a twice-differentiable function on $[x_1, x_n]$ and λ is the smoothing parameter indicating the roughness of the resulting spline function. Cubic splines for given $x_1 < t_1 < t_2 < ... < t_n < x_n$ are cubic polynomials on each interval $[x_1, t_1], [t_1, t_2], ..., [t_n, x_n]$ and the polynomial pieces fit together at points t_i, meaning $\hat{f}$ itself and its first and second derivatives are continuous at each t_i, and hence on the whole $[x_1, x_n]$. The cubic splines are specified as

$$\hat{f}(t) = d_i(t - t_i)^3 + c_i(t - t_i)^2 + b_i(t - t_i) + a_i \tag{3.2.5}$$

for $t_i \leq t \leq t_{i+1}$.

As this method intends to estimate sample sizes, a maximum value x_{max} of this function is predicted at $2 * x_n$ and then the fitted values $\hat{f}(0.9 * x_{max})$ as well as $\hat{f}(0.95 * x_{max})$ are calculated. The R package stats is used to fits a cubic smoothing spline to the supplied data and to predict the smoothing spline fit at new points.

3.2.8. Functional analysis of signatures

For biological interpretation of gene signatures, we performed an enrichment analysis of gene ontology (GO) terms (*Ashburner et al.*, 2000) using GOrilla (*Eden et al.*, 2009) as well as a cancer-related disease gene enrichment analysis using FunDO (*Du* *et al.*, 2009). GOrilla discovers enriched GO terms in a target set versus a background set (e.g. all transcripts present on the microarray). FunDO uses the Disease Ontology annotation (*Osborne et al.*, 2009) to provide analysis of disease terms associated with genes in a gene list. Both tools evaluate significance of associations by a Fisher's exact test based on hypergeometric distribution.

3.2.9. Implementation of a web-based AML diagnosis tool

A tool for AML classification was implemented using RWui (*Newton et al.*, 2011). This tool performs AML diagnosis based on the entire AML HG-U133A data set. A SVM classifier is established on 200 best performing features and can be applied to a new HG-U 133A sample uploaded by the user.

4. Establishment of the conventional approach for biomarker development

In this chapter a conventional microarray classification workflow is established for the identification of non-small cell lung cancer (NSCLC) in comparison to control samples. This workflow comprises

- Comparison of feature selection cut-offs and classification algorithms in a training cohort by CV
- Building an optimized classifier in the TS using the chosen features and algorithm
- Application of this classifier to two independent validation cohorts and assessment of classifier performance

This approach is further applied to one of the gene expression datasets, a multiple myeloma dataset, provided by the MAQC consortium (*Shi et al.*, 2010).

4.1. Motivation

Lung cancer is still the leading cause of cancer related death worldwide. Prognosis has remained poor with a disastrous five year survival rate of only about 15% due to diagnosis of the disease in late, i.e. incurable stages in the majority of patients (*Jemal et al.*, 2008) and still disappointing therapeutic regimens in advanced disease (*Sandler et al.*, 2006). Thus, there is an urgent need to establish reliable tools for the identification of NSCLC patients at early stages of the disease e.g. prior to the development of clinical symptoms. Today, the only way to detect non-small cell lung cancer is by means of imaging technologies detecting morphological changes in the lung in combination with biopsy specimens taken for histological examination. However, these screening approaches are not easily applied to secondary prevention of non-small cell lung cancer in an asymptomatic population (*Henschke et al.*, 2006). The use of surrogate tissue-based, e.g. blood-based, biomarkers for non-small cell lung cancer might therefore circumvent the known pitfalls of imaging technologies and invasive diagnostics (*Henschke et al.*, 2006; *Bach et al.*, 2007). Such biomarkers might be utilized to direct imaging based and invasive screening approaches to only those individuals identified as potential non-small cell lung cancer patients by biomarker screening. Array-based assessment of disease-specific gene expression patterns in peripheral blood mononuclear cells (PBMC) have been reported for non-malignant (*Staratschek-Jox et al.*, 2009) and malignant diseases including renal cell carcinoma, melanoma, bladder, breast and lung cancer (*Burczynski et al.*, 2005; *Twine et al.*, 2003; *Sharma et al.*, 2005; *Osman et al.*, 2006; *Critchley-Thorne et al.*, 2007; *Showe et al.*, 2009). In some cases gene expression profiles derived from PBMC were even suggested as promising tools for early detection (*Showe et al.*, 2009; *Sharma et al.*, 2005) or prediction of prognosis (*Burczynski et al.*, 2005), albeit these findings have not yet been validated in independent studies. Furthermore, circumventing known pitfalls of analyzing PBMC in a clinical setting (*Debey et al.*, 2006, 2004) by using stabilized RNA derived from whole blood would further strengthen the validity of blood-based surrogate biomarkers for early diagnosis of lung cancer and other malignant diseases. We therefore investigated the validity of whole blood-based gene expression profiling for the detection of NSCLC patients among hospital based controls as well as healthy individuals.

4.2. Results

4.2.1. Establishment of a gene expression profiling-based classifier for blood-based diagnosis of NSCLC

The classifier was build based on an initial TS containing 35 NSCLC cases of different stages (stage I: n=5, stage II: n=5, stage III: n=17, stage IV: n=8) and 42 hospital based controls suffering in part from severe comorbidities such as COPD, hypertension, cardiac diseases as well as malignancies other than lung cancer (see Tab. 4.1).

We first evaluated three different approaches, namely SVM, LDA and PAM to identify the best algorithm to build a classifier for the diagnosis of NSCLC in a 10-fold CV design (see Fig. 4.2.1A). To this end we used 36 different feature lists extracted from the list of differentially expressed genes according to 36 different cut-off p-values of the T-statistics. In this setting the SVM algorithm performed best by reaching the highest AUC (mean AUC = 0.754) at a cut-off p-value of the T-statistics of 0.003 (Fig. 4.2.1B). Thus for subsequent classification we applied SVM by using the 484 feature list obtained at a cut-off p-value of the T-statistics of p-value<0.003 for differential expressed genes between cases and controls based on the entire TS.

	Training set (TS)		Validation set (VS1)		Validation set (VS2)	
	NSCLC	controls*	NSCLC	controls*	NSCLC	controls**
total number	35	42	28	26	32	70
female	10	14	9	10	16	35
male	25	28	19	16	16	35
median age	61	61	62	65	67	44
NSCLC stage 1	5	NA	6	NA	32	NA
NSCLC stage 2	5		2		0	
NSCLC stage 3	17		12		0	
NSCLC stage 4	8		8		0	

Table 4.1.: Clinical and epidemiological characteristics of cases with lung cancer and respective controls
Clinical and epidemiological characteristics of cases and controls in the training set as well as the two independent validation sets are given. *Hospital based ** healthy blood donors; median age is given in years, NA: not applicable

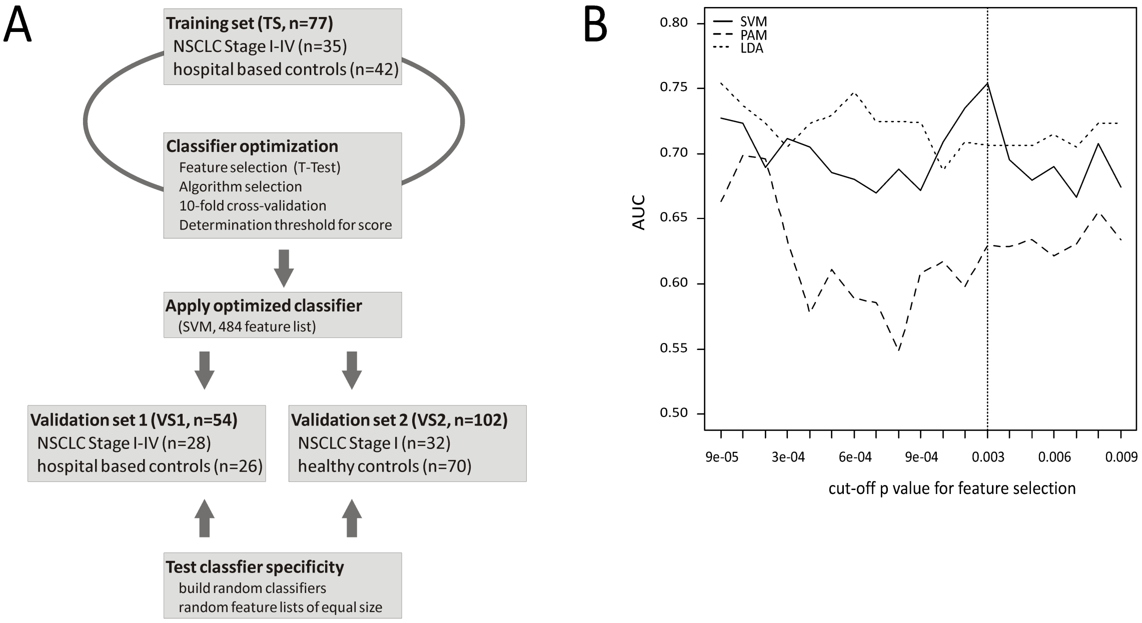


Figure 4.2.1.: Experimental design and parameter optimization
A: In the training set (TS) the optimal classifier was established, and then applied to two validation sets VS1 and VS2. To test the specificity of this optimized classifier additional 1,000 classifiers using random feature lists of equal size were permuted and applied to VS1 and VS2. The workflow layout was originally designed by Dr. T. Zander. B: Identification of the optimal algorithm for classification based on the TS. The AUC is plotted against the cut-off p-value of the T-statistics for feature selection for all three algorithms (SVM, LDA, PAM) in the 10-fold cross-validation of the TS. SVM leads to the highest mean AUC (mean AUC = 0.754) at a cut-off p-value of 0.003 which is highlighted by a dotted line.

4.2.2. The diagnostic NSCLC classifier can be used to detect NSCLC cases in two independent validation sets

First, we validated whether the classifier can be used to discriminate NSCLC cases of early and advanced stages among hospital based controls. Therefore in the first independent VS, cases and controls were chosen in a similar setting as in the TS, this is, patients with NSCLC stage I to IV and clinical symptoms associated with lung cancer and hospital based controls with relevant comorbidities (n=26). The AUC for the diagnostic test of NSCLC in this first VS (VS1) was calculated to be 0.824 (p-value<0.001, Fig. 4.2.2A). In addition, we observed a significant difference between the SVM based probability scores to be a NSCLC case for actual NSCLC cases and controls in VS1 (p-value<0.001, T-test).

After demonstrating that the classifier can be used to detect NSCLC cases among individuals with comorbidities we also investigated whether this test can be used to distinguish NSCLC cases presenting at stage I with no or only minor symptoms from healthy individuals. Therefore we recruited a second independent VS consisting of 32 NSCLC cases at stage I and 70 healthy blood donors (VS2). By applying the identical classifier to VS2 the AUC was determined to be 0.977 (p-value<0.001, Fig. 4.2.2B). We also observed a highly significant difference in the probability values to be a NSCLC patient for cases in contrast to controls (p-value<0.001, T-test). Healthy controls without significant comorbidity (VS2) tend to have lower probability scores compared with hospitalbased controls (VS1 and TS) although this finding was not statistically significant (Fig. 4.2.2D).

4.2.3. Permutation test to analyse the specificity of the classifier

To further underline the specificity of this classifier, we used 1000 random feature lists each comprising 484 features to likewise build a SVM-based classifier in the TS which then were applied to VS1 and VS2, respectively (Fig. 4.2.2C). For VS1 the mean AUC obtained by using these random feature lists was 0.6839 with 31 AUCs being $\leq$ 0.824, the AUC obtained using the NSCLC specific classifier. This corresponds to a p-value of 0.031 for the permutation test further confirming the specificity of the NSCLC classifier. Similar by applying the permuted classifiers to VS2 only 11 of random feature lists lead to an AUC of $\leq$ 0.977, the AUC obtained using the NSCLC-specific classifier (p-value=0.011). In conclusion, a NSCLC-specific blood-based classifier was build that was successfully used to identify NSCLC cases among hospital based controls as well as NSCLC cases of early stage among healthy individuals.

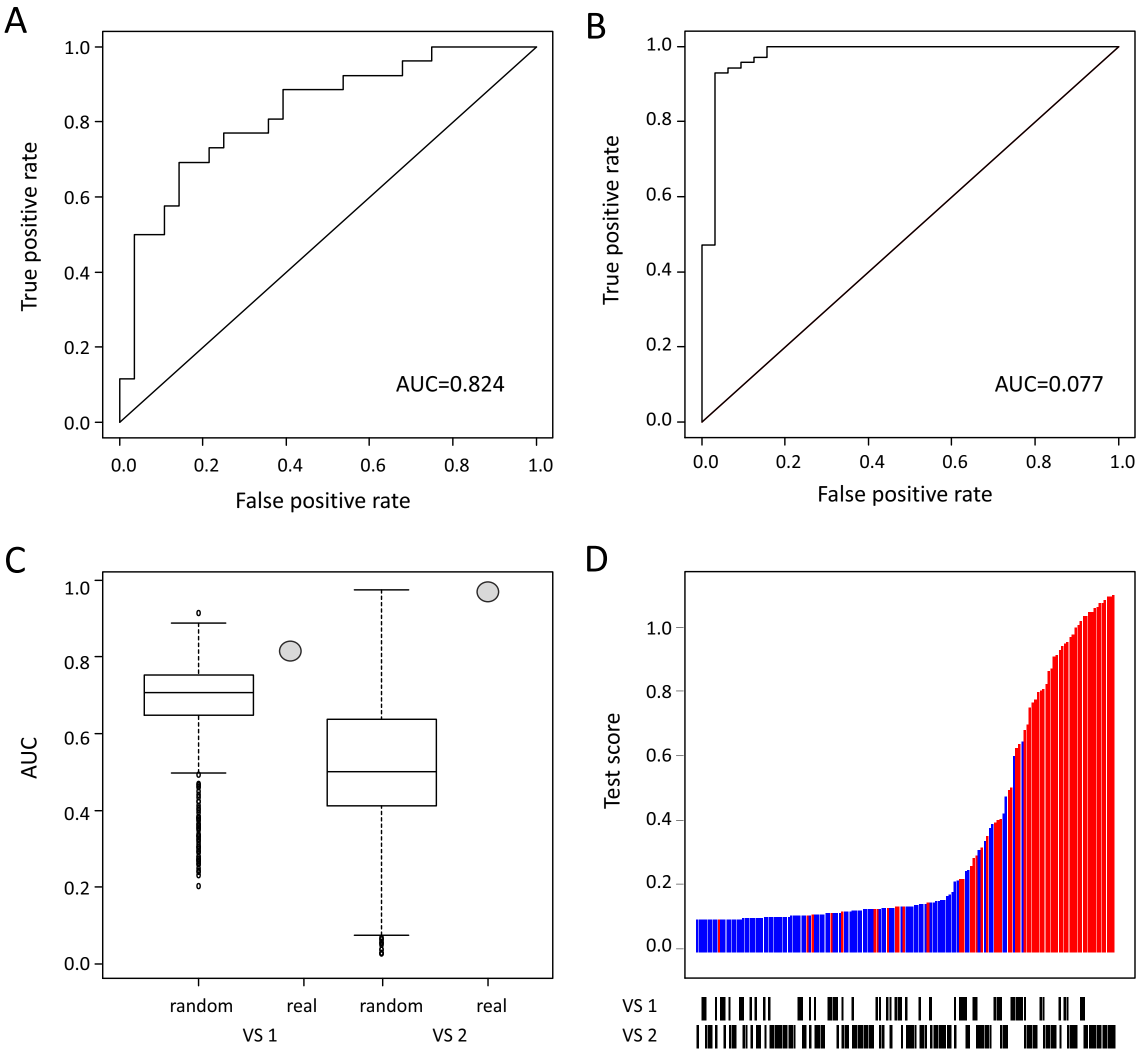


Figure 4.2.2.: Performance of optimized classifier in VS1 and VS2
(A) Receiver operating characteristic (ROC) curve for the optimized classifier established in the training set applied to validation set 1 (VS1: all stage NSCLC patients and hospital based controls), AUC = 0.824. (B) ROC curve for the optimized classifier applied to validation set 2 (VS2: stage I NSCLC patients, healthy controls), AUC = 0.977. Shown is the true positive rate (Sensitivity) on the y-axis against the false positive rate (1-Specificity) on the x-axis. (C) Box plots comprising 1,000 AUCs obtained by using a random list of 484 features to build the classifier in TS and then apply it to VS1 (left) and VS2 (right). The real AUC using the specific classifier is depicted with grey circles. (D) Test scores to be a case of all samples from VS1 and VS2 were ranked. NSCLC cases are marked in red and controls in blue. Membership in a specific cohort is indicated by a vertical line underneath the graph. For visualization of the test score obtained by the SVM algorithm we used the following transformation: $log_2(score + 1) + 0.1$. The layout was adapted from a graphic originally designed by Dr. T. Zander.

4.3. Summary and Discussion

Using RNA-stabilized whole blood from smokers in three independent sets of NSCLC patients and controls we present a gene expression based classifier that can be used as a biomarker to discriminate between NSCLC cases and controls. The optimal parameters of this classifier were first determined by applying a classical 10-fold CV approach to a TS consisting of NSCLC patients (stage I-IV) and hospital based controls (TS). Subsequently this optimized classifier was successfully applied to two independent VSs, namely VS1 comprising NSCLC patients of stage I-IV and hospital based controls (VS1) and VS2 containing patients with stage I NSCLC and healthy blood donors. This successful application of the classifier in both VSs underlines the validity and robustness of the classifier. Extensive permutation analysis using random feature lists and the possibility of building specific classifiers independently of the composition of the initial TS further support the specificity of the classifier. We found no associations between stage of disease and the probability score assigned to each sample. In addition we observed no association between other cancers and the probability score of the controls (data not shown). But controls without documented morbidity (controls in VS2) tend to have lower probability scores to be a case as compared to controls with documented morbidity, although this was not statistically significant.

Recently, Showe et al. (*Showe et al.*, 2009) reported a NSCLC associated gene expression signature derived from PBMC of predominantly early stage NSCLC patients. Since we used RNA-stabilized whole blood and not PBMC for analysis, we were not surprised that the signature identified by Showe et al. could not be used in our dataset to distinguish between cases and controls. The same holds true when applying our classifier to the published data set (data not shown). Findings derived from several of our own studies further underline that signatures derived from PBMC and RNA-stabilized whole blood samples cannot be directly compared (*Debey et al.*, 2006, 2004). However, as previously shown by us and others, for clinical applicability and robustness we would favor RNA-stabilized approaches since these methods reveal more reliable results in a multi-center setting (*Debey-Pascher et al.*, 2011).

Overall, our data demonstrate the feasibility of a diagnostic test for NSCLC based on RNA-stabilized whole blood. Our findings form the basis for validation studies in a multi-center setting in prevalent NSCLC patient cohorts enriched for early stage disease. At the end, this endeavor might open the avenue to test the blood-based NSCLC classifier in prospective trials to evaluate the predictive potential of diagnostic classifiers for NSCLC in high-risk individuals.

4.4. Further analysis of MAQC-II data

To evaluate whether our approach is comparable to standards defined in the MAQC-II study (*Shi et al.*, 2010), we applied our preferred method established before for data processing, feature selection and classifier development to one of the 6 GEX data sets provided by the MAQC consortium. This multiple myeloma data set (*Zhan et al.*, 2006; *Shaughnessy et al.*, 2007) comprised a total of 340 samples in the TS plus 214 samples in the VS and defined four different endpoints, namely event free survival (EFS) meaning a lack of malignancy or disease recurrence, overall survival (OS) after 730 days, gender (representing the sex of patients, which is highly predictable by the microarray data; positive control), and random (randomly assigned sample class labels that are not predictable by microarray data; negative control) measured on the Affymetrix HG-U133A microarray. We defined a model (FC/p-value filter for feature selection, linear SVM classifier) without further optimization by internal 10-fold CV which was repeated 10-times in the TS (internal validation) and applied this model to the independent VS (external validation). These results were compared to classification performance of best-performing models defined by MAQC-II data analysis teams for all 4 endpoints.

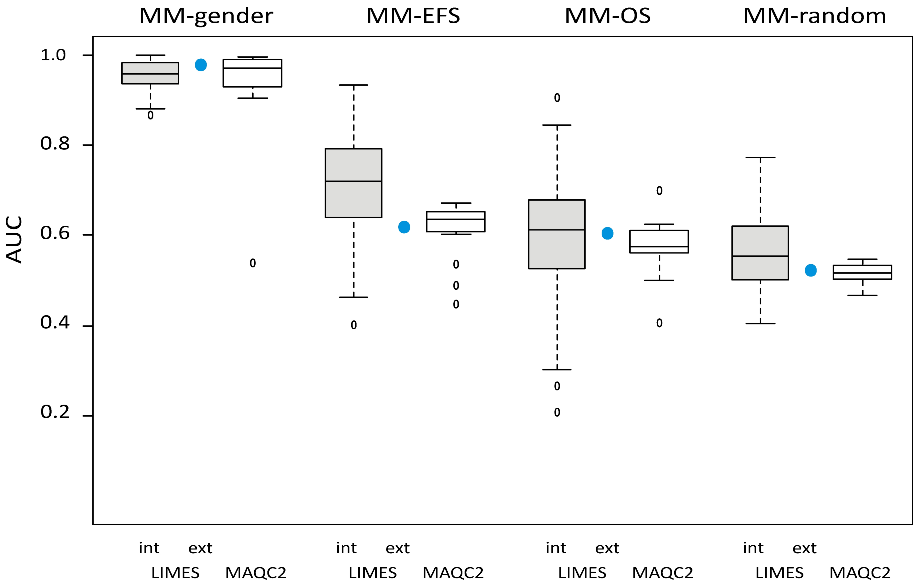


Figure 4.4.1.: Distribution of AUC performance of our classifiers (LIMES) defined by internal 10-fold CV repeated 10-times (int, grey), external validation (ext, blue) as well as the best-performing models of MAQC-II (n = 17, white). Endpoints event free survival (EFS), overall survival (OS), gender (positive control), and random (randomly assigned sample class labels, negative control) of the multiple myeloma dataset provided by MAQC-II were assessed. Boxes indicate the 25% and 75% percentiles, and whiskers indicate the 5% and 95% percentiles. The layout was adapted from a graphic originally designed by Prof. Dr. J. L. Schultze.

As demonstrated by Fig. 4.4.1, both internal and external classification performance results for AUC across all 4 endpoints are similarly well to best-performing models obtained by data analysis teams included in the MAQC-II. Similar results were obtained for MCC, sens and spec (data not shown). Overall, our analysis approach to make clinical predictions is up to the MAQC-II standards and good modeling practice is provided.

5. Simulation and adaptive learning approaches enhance classification

This chapter introduces methods to enhance molecular diagnostics including novel strategies to predict the outcome of a large biomarker study from a small pilot study and to estimate approximate sample size of future validation cohorts. These concepts are developed on a large acute myeloid leukemia (AML) gene expression data set and then applied to predict the diagnosis of NSCLC, active tuberculosis (ATB) and human immunodeficiency virus (HIV) from microarray data.

5.1. Motivation

Irrespective of widespread establishment of transcriptional-based biomarkers, the number of gene signatures that have entered clinical practice is alarmingly small and the translation of basic findings to clinical utility such as diagnosis and prognosis has been slow. As already pointed out in the introduction, the MAQC-I study (*Shi et al.*, 2006) has successfully demonstrated that the microarray technology itself is reliable and reproducible. As questions remained regarding the reliability of the technology in clinical applications such as disease diagnostics or prognostics, the MAQC-II project (*Shi et al.*, 2010) was launched to investigate the capabilities and limitations of microarray technology. Various data analysis methods in developing and validating of predictive signatures were assessed and classifier models and a consensus on the "best practices" for these models was reached. Important results from this large consortium effort include an outcome dependent model prediction performance and a high degree of concordance between well-implemented internal and external validation. As many classifiers with similar statistical performance can be identified for a studied endpoint, a great uncertainty about selecting the "optimized" model remains.

In this project we sought to address several unresolved issues. First, to better judge the validity of small pilot trials we developed a two-step validation approach combined with randomized permutation ("clinical trial simulation"). Second, to predict the minimum size of a consecutive pivotal validation trial we describe an algorithm combining sample simulation and adaptive learning approaches ("on the fly optimization strategy"). This approach can also estimate overall best test performance. Utilizing these approaches we introduce a high-performance test for primary molecular diagnosis of leukemia, we estimate the trial size and test performance for pivotal trials using peripheral blood to develop three independent biomarker tests to detect patients with ATB, NSCLC and HIV.

5.2. Results

5.2.1. Establishment of AML GEX data as a reference set

The evaluation of simulation and adaptive learning approaches to accelerate clinical biomarker development first required the establishment of a sufficiently large set of high-throughput data suitable for assessing a clinically relevant endpoint. Ideally, the data set should be derived from a common data space based on a single technology. This feature would be particularly beneficial for normalization issues of data derived from multiple locations and/or studies. Therefore, we searched for a diagnostic setting that would fulfill the following criteria: 1) sufficient data on a single technical platform, even if from different sources, 2) ability to integrate such data with high enough quality, 3) prior knowledge about the performance of such technology in a diagnostic setting, 4) a disease setting with a high enough prevalence, and 5) the need to further improve molecular diagnostics to improved therapy and disease outcome. High dimensionality of GEX data and recently suggested guidelines by the MAQC consortium made this technology a prime source to test our overall approach.

We chose primary molecular diagnosis for AML as the first model endpoint. Several studies already reported the successful establishment of transcriptional-based classifier for disease subclassification, outcome prediction and differential diagnosis (*Bacher et al.*, 2009a,b). Surprisingly, primary molecular diagnosis was not an endpoint in these studies. AML is a malignant disease of the myeloid white blood cell line characterized by a rapid growth of abnormal white blood cells. Malignant cells interfere with normal blood cell production and initially accumulate in bone marrow and peripheral blood with further suppression of its normal counterparts and later infiltrate solid organs and tissues. Patients with AML typically have a poor prognosis. Importantly, current diagnostic strategies involve bone marrow examination by an experienced hematologist, flow cytometry, PCR as well as cytogenetic analysis (*Cheson et al.*, 2003). This rather time-consuming procedure does not allow for designing novel induction therapies according to risk factors. Thus, there is obvious need for diagnostic tools to rapidly confirm primary diagnosis by light microscopy while at the same time ascertain prognostic and predictive factors. We therefore investigated the use of RNA microarray analysis of peripheral blood or bone marrow as a tool for primary diagnosis that might substitute currently used diagnostic technologies. As AML has a low prevalence but is a deadly disease if not diagnosed in time, a test used for screening or primary diagnosis of AML would have to achieve sensitivity and specificity $> 97\%$ preferably $>99\%$ to minimize false-negative results while avoiding unacceptable levels of false-positive results.

Positives	Negatives	Further sample description	Reference
43		AML	*Gutiérrez et al.* (2005)
50		AML	*Haferlach et al.* (2005)
	50	ALL	
	98	CML and CLL	
163		AML	*Metzeler et al.* (2008)
150		AML	*Ross et al.* (2004)
	5	ALL	
26		AML	*Stirewalt et al.* (2008)
	38	Healthy controls	
285		AML	*Valk et al.* (2004)
	5	Healthy controls	
	175	ALL	*Mullighan et al.* (2009)
	108	Healthy controls	*Baty et al.* (2006)
	14	Healthy controls	*Borovecki et al.* (2005)
	17	Huntington's disease	
	42	Healthy controls	*Burczynski and Dorner* (2006)
	85	UC/ Crohn's disease	
	8	Healthy controls	*Ramilo et al.* (2007)
	262	Infectious diseases	
	14	Healthy controls	*Connolly et al.* (2004)
	26	Healthy controls	Debey-Pascher
	7	Healthy controls	*Gow et al.* (2009)
	8	CFS	
	121	Healthy controls	*Sinnaeve et al.* (2009)
	101	Coronary artery disease	
	20	Healthy controls	*Scherzer et al.* (2007)
	85	Parkinson disease	
	7	Healthy controls	*Watford et al.* (2008)
717	**1296**	$\sum$	

Table 5.1.: HG-U133A AML data set
Summary of GEX studies included in the HG-U133A AML data set. Numbers shown are the actual number of samples used in the final data set. In total, this data set includes 2013 samples from 17 different studies.

In an initial screen we identified 113 studies addressing diagnostics issues in leukemia and for only 45 studies transcriptome data were available. Seven of these studies were performed on a single microarray platform using the Human Genome model HG-U133A microarray from Affymetrix. Only one of these studies addressed molecular primary diagnosis of AML as the major endpoint; however, this study was extremely small with only 26 AML patients and 38 healthy controls (*Stirewalt et al.*, 2008). All other studies focused on differential diagnosis respectively subclassification or disease and treatment outcome (*Valk et al.*, 2004; *Ross et al.*, 2004; *Haferlach et al.*, 2005; *Gutiérrez et al.*, 2005; *Metzeler et al.*, 2008). We therefore integrated additional data sets to be able to address the important question of AML primary diagnosis with sufficiently high statistical power. In total, 2013 samples from 17 different studies passing defined quality control criteria (see chapter 3) were compiled to form a new data set (Tab. 5.1). These studies included samples from AML (n=717) and healthy controls (n=410) as well as samples comprising other diseases not related to leukemia (n=558) and other leukemia samples including acute lymphoblastic leukemia (n=230) and chronic leukemia (n=98).

5.2.2. Trial simulation approach

5.2.2.1. Performance of a small pilot trial

To simulate a typical pilot trial study (PTS) setting as a first step of test development, 150 samples were drawn randomly from the complete data set and distributed into the three sets TS, VS1 and VS2, each containing 25 AML cases and 25 controls (Fig. 5.2.1A). This initial setting allowed the development of a classifier within TS and two independent validations (in VS1 resp. VS2). Since classifier performance in subsequent validation cohorts might be strongly influenced by the characteristics of the patient population within TS, it was already suggested in MAQC-II to perform swap analysis of training and validation cohorts (*Shi et al.*, 2010). In principle, swapping independent patient cohorts is just a special case of random permutation of samples. We therefore extended this approach from a single swap to 10,000 permutations (termed '10,000 trial simulation approach'; TSA) containing for each simulation random drawing of one TS and two independent VSs (Fig. 5.2.1A). Using our conventional linear SVM-based classification with a FC/p-value filter (FC 10, p-value 0.0001) for feature selection, a relatively high mean AUC of 0.9688 in both VSs was achieved by TSA for this AML pilot trial (Fig. 5.2.1B). Taken together, TSA allows quantification of the influence of sample distribution (into the respective cohorts) on the statistical performance of the classifier and average metrics gives an early estimate for overall classifier performance to be expected for the endpoint under study.

Next classification algorithms, feature selection, or size and sample distribution within TS were varied to elucidate their influence on prediction performance. Again, AUC, MCC, sensitivity and specificity were evaluated by TSA generating a total of 210,000 classifier. We first used a different SVM kernel and further compared SVM to LDA and PAM algorithms (n= 8×10,000 classifier, Fig. 5.2.1C). SVM and PAM performed similarly well while LDA did not reach comparable high mean AUC. The non-parametric wilcoxon test resulted in more features not passing the filter cut-offs.

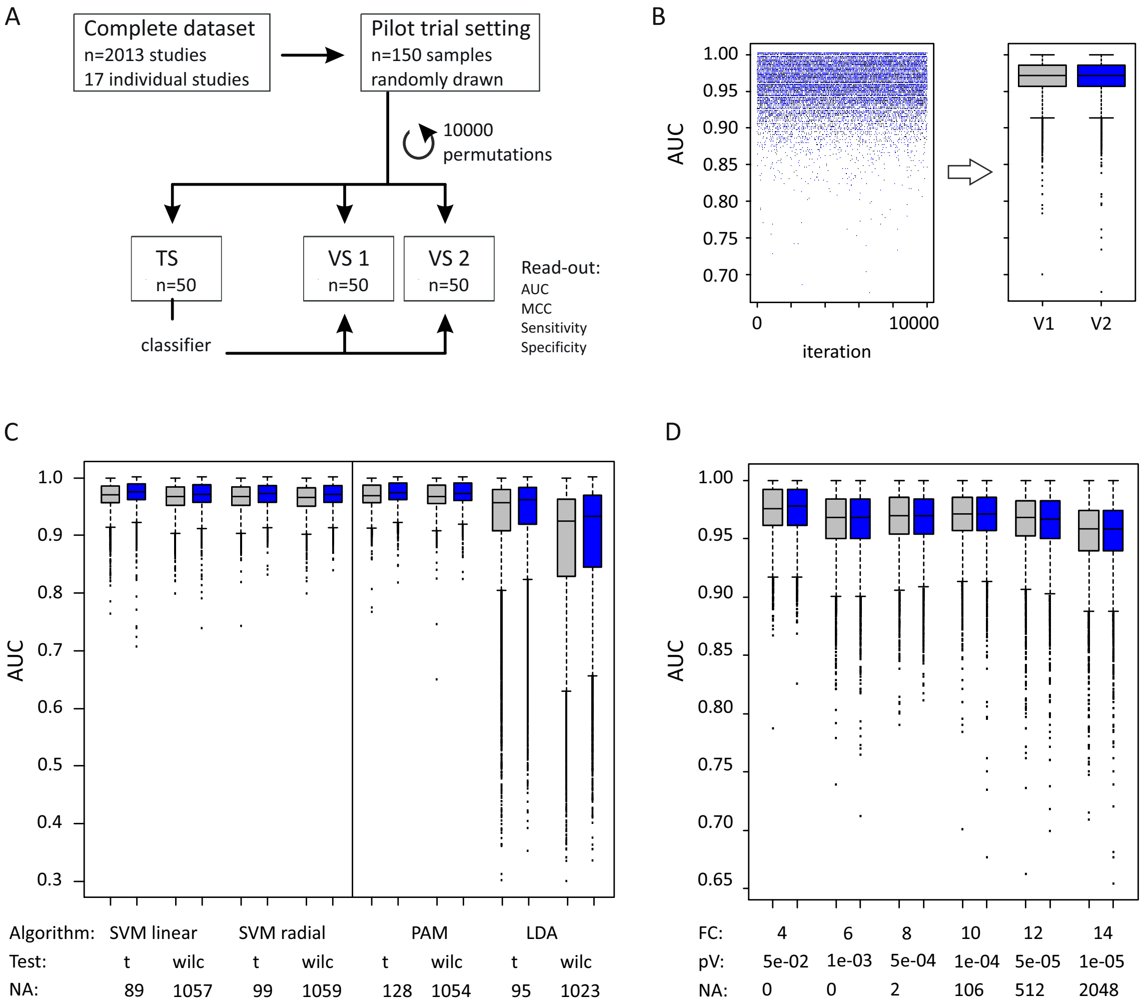


Figure 5.2.1.: Classification performance of AML pilot trial
(A) The pilot trial study simulation for AML is shown as a schematic view. First, a small pilot trial study (PTS) is drawn from the complete cohort. This PTS is divided in a training set (TS) and two validation sets (VS1, VS2) without replacement. Feature selection and classifier construction is based on the TS. The TS classifier is applied to the 2 VSs. The entire process is repeated 10,000 times. The workflow layout was adapted from a graphic originally designed by Prof. Dr. J. L. Schultze. (B) AUC values for 10,000 iterations of classification, displayed are individual classification results (left panel) for VS1 (grey) and VS2 (blue) and the summarizing boxplots (right panel). Boxes indicate 25th and 75th percentiles, the line within the box marks the median, whiskers above and below the box indicate 90th and 10th percentiles, outliers are plotted as dots. Distribution of AUC performance of (C) different classifier and feature selection algorithms (SVM with linear and radial kernel, PAM and LDA) in combination with a t-test or Wilcoxon test and (D) different feature size cut-offs for fold-change (FC)/p-value (pV) filter (D) are displayed as boxplots. NA values indicate the number of tests without features passing the filter cut-offs.

Since TSA clearly established the framework for classifier performance (range, median, 75% percentile) we next addressed dependency of classifier performance on feature size and feature selection cut-off (n= 5×10,000 classifier, Fig. 5.2.1D). Unexpectedly, reducing feature size resulted in inferior classifier performance suggesting that due to the overall small sample size, more features are required to correctly classify in independent validation cohorts (Fig. 5.2.1D). This was similarly true when reading out MCC, sensitivity and specificity (Fig. B.0.1).

For completion, TSA also clearly established that further reduction of sample size or unequal distribution of samples in TS results in reduced overall classifier performance (Tab. 5.2).

	%AML	%Control	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
VS1	25	25	0.728	0.9568	0.9712	0.9683	0.9856	1
	20	25	0.7248	0.952	0.968	0.9644	0.984	1
	15	25	0.632	0.9424	0.9632	0.958	0.9808	1
	10	25	0.6752	0.9344	0.9584	0.9515	0.9776	1
	5	25	0.1008	0.9184	0.9504	0.9383	0.9728	1
VS2	25	25	0.7648	0.9568	0.9712	0.9686	0.9856	1
	20	25	0.7696	0.952	0.968	0.9645	0.984	1
	15	25	0.6672	0.944	0.9632	0.9582	0.9808	1
	10	25	0.6448	0.9344	0.9584	0.9514	0.9776	1
	5	25	0.0928	0.9184	0.9504	0.9387	0.9728	1
VS1	20	20	0.648	0.9504	0.968	0.9634	0.9824	1
	15	15	0.6992	0.9376	0.96	0.9541	0.9776	1
	10	10	0.6128	0.92	0.9488	0.9392	0.9696	1
	5	5	0.3648	0.8832	0.9312	0.9118	0.96	1
VS2	20	20	0.692	0.9504	0.9664	0.9635	0.9824	1
	15	15	0.6448	0.9376	0.96	0.9543	0.9776	1
	10	10	0.5936	0.92	0.948	0.9391	0.9696	1
	5	5	0.408	0.8824	0.9312	0.9114	0.9616	1

Table 5.2.: AUC performances for different case:control proportions
Shown are the summary statistics of AUC performance for different proportions of AML to control samples in 10,000 permutations of TSA.

Overall, TSA seems to be well-suited to establish overall classifier performance that can be expected independent of the actual clinical situation with subsequent patient recruitment into TS and validation cohorts for the endpoint under study. Moreover, dependencies of classifier performance and the influence of sample distribution, classification algorithms and feature selection are easily uncovered.

Depending on the sample distribution to the three independent cohorts (TS, VS1, VS2), performance of a significant number of classifiers would not have supported further classifier development. Moreover, many classifier did not reach the set target of performance (e.g. AUC> 0.99). To further elucidate whether one single drawing of samples into the three cohorts predicts the outcome in a large validation cohort, we reanalyzed all classifiers developed during feature cut-off filter variation (n=60,000).

First, we estimated the number of classifiers who would have been candidates to be further used in clinical practice by the proportion of classifiers in TSA with mean AUC in VS1 and VS2 >0.99. In total, 1451 out of 60,000 classifiers (2.4183%) fulfilled this criterion. We next applied those classifiers to the remaining samples from the large AML cohort (n=1863). Only 572 classifiers (39.4211%) from the candidate classifiers resulted in a AUC value >0.99 for this larger validation cohort and vice versa, 60,5789% would result in inferior prediction performance. AUC distributions of these classifiers are visualized in the density plot of Fig. 5.2.2 and correlation between internal and external validation in this setting are shown in Fig. B.0.2. These results further demonstrate that the classical design of fixed sample drawing into the different cohorts is an inadequate approach to establish a final classifier.

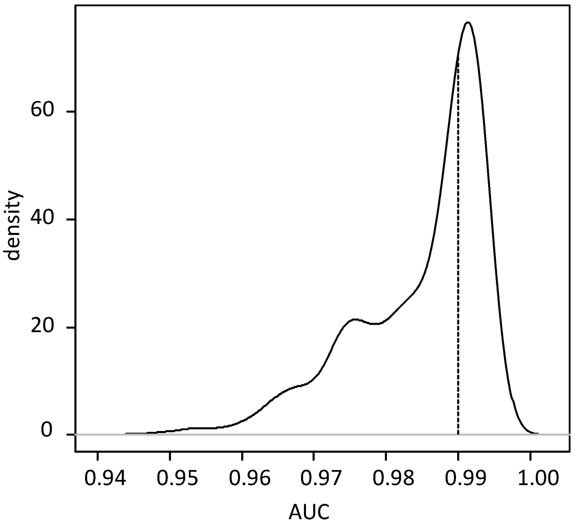


Figure 5.2.2.: Distribution of AUC values of external validation for classifiers resulting in AUC values >0.99 in internal validation, displayed as a density plot. A dashed horizontal line indicate AUCs >0.99.

5.2.2.2. Selection of a good classifier from the pilot study

There are several ways to select a good classifier from a pilot trial. Usually, this classifier will then be fixed and validated in a big VS. Conservative approaches split the PTS into a pilot TS and a pilot VS, A single classifier is trained on the pilot TS and evaluated on the pilot VS. A major drawback of this approach is that only 50% of the data is used for classifier training. As already shown in Fig. 5.2.2, a high proportion of classifiers would result in overoptimistic performance results.

Our approach is building from the whole PTS using features identified in the TSA and in this way, expecting to get a robust feature ranking from different sample subsets. We choose four different approaches to identify and rank informative features in TSA (see chapter 3). First, we counted the number of times a feature is identified as differentially expressed in each TSA iteration and each feature can then be ranked according to this criteria (“**DE selection**”). By using subsets of the original data set, stable markers would be expected to appear more often than uninformative features in the signatures. Second, we assessed features together with classification performance using the score assigned to each feature. In linear SVM classification, the absolute values of the weights of each feature can be regarded as the contribution of each feature and a ranking can be obtained according to the weighted ranksum (**“weights SVM**”). Both methods focuses on DE features. As other features not chosen by this strict filter might be informative as well, we also calculated the ranksum of all features according to both the calculated p-values (“**pV ranking**”) and FCs (“**FC ranking**”). A comparison of feature ranking revealed that DE ranking and weighted SVM resulted in comparable rankings, with similar features identified as best performing features. In comparison, pV and FC rankings let to a higher variance in ranking comparison and pV ranking was the least comparable method regarded to specific feature ranks. All rankings are visualized in Fig. B.0.4.

All features identified during 10,000 TSA iterations are saved and summarized. Then a classifier is build in the PTS AML cohort of 150 samples using the top 10, 20, 50, 100 and 200 features identified in all four approaches, namely DE selection, weights SVM, pV and FC ranking. These 20 classifiers are then applied to the remaining samples from the large AML cohort (n=1863). As shown in Fig. 5.2.3A and Fig. 5.2.3B, all rankings let to stable classification results with nearly similar performance (mean AUC = 0.9765). Feature ranking using the weights of a linear SVM resulted in slightly higher AUC values for 3 out of 5 classifiers and in higher MCC values for 4 out of 5 classifiers. Minimal performance was obtained when using features ranked according to a p-value obtained by statistical testing (“pV ranking”). For completion, we also calculated the 10, 20, 50, 100 and 200 best performing features from all 150 samples in the PTS without any subsampling. As demonstrated by the grey barplots of Fig. 5.2.3A and Fig. 5.2.3B, feature ranking by weighted SVMs also outperformed the performance obtained without subsampling. All five classifiers using the weights SVM method were also significantly superior to random classifiers of equal size as tested in permutation analysis (Fig. 5.2.3C). Hence, the weights SVM feature ranking method was further used to build final diagnostic signatures in the other data sets.

5.2.2.3. Comparison of trial simulation approach to bootstrap design

Additionally, we compared our TSA (sample-split without replacement) approach to a classical bootstrap design where samples are drawn into TS and the two VSs with replacement. When analyzing the prediction performance of the resampling

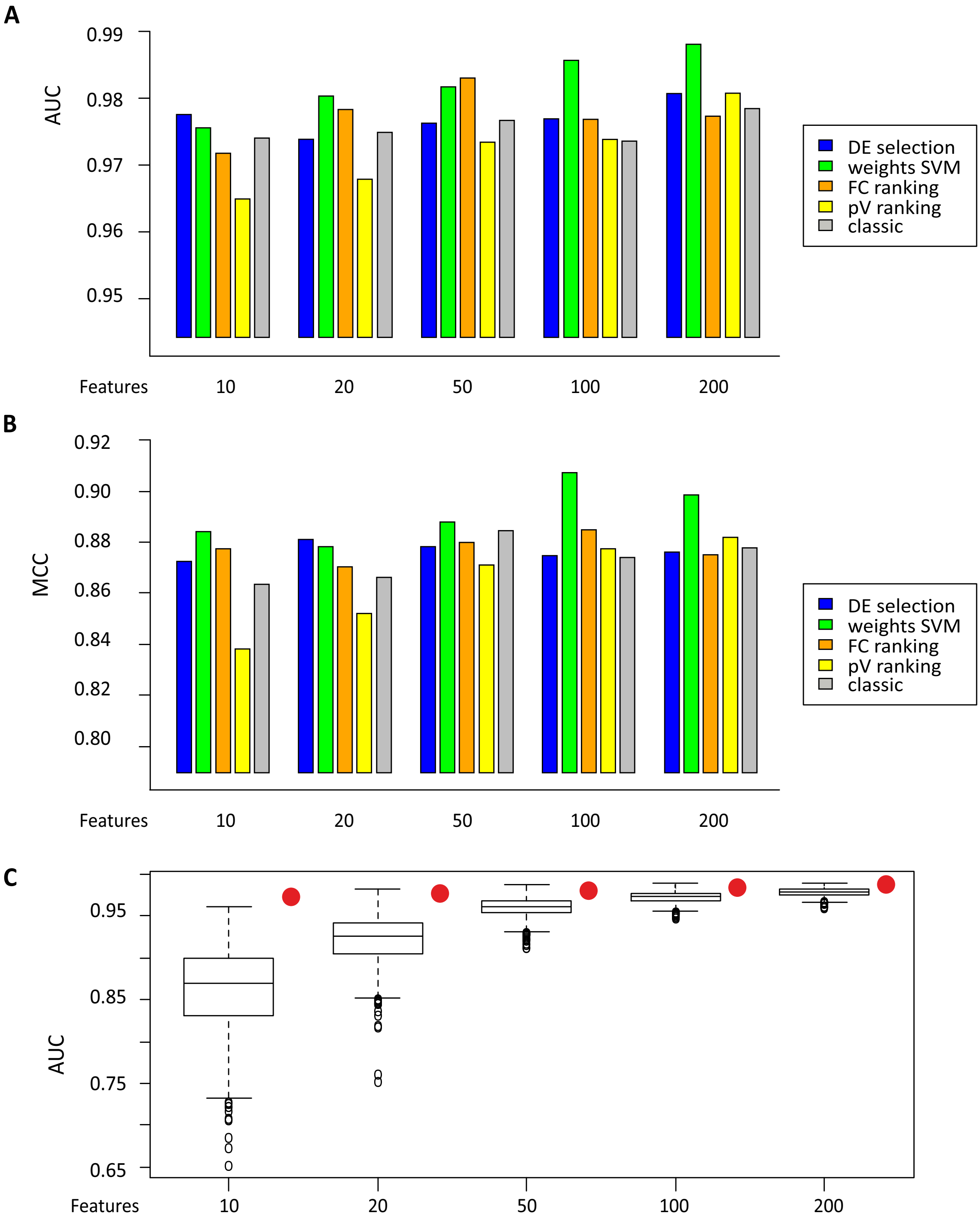


Figure 5.2.3.: Selection of a good classifier in pilot AML cohort
A-B: Shown are AUC (upper panel) and MCC (lower panel) prediction results for both all four different feature ranking methods obtained in 10,000 iterations of TSA for the small pilot AML trial and a classical approach without subsampling using the best ranked 10, 20, 50, 100 and 200 features applied as a classifier to the remaining samples of the large AML cohort. C: Box plots comprising 1000 AUCs obtained by using random feature lists of equal size. The real AUC using the specific weights SVM classifier is depicted with red circles.

designs for different FC/p-value filter cut-offs, the classical bootstrap design leads to significantly higher AUC values in comparison to our approach (p-value $< 10^{-16}$ for all FC/p-value filter, see density plots (sec. 2.3.1) in Fig. 5.2.4 and Fig. B.0.3) as expected due to the overlap of samples between TS and VS. The same results are obtained for other performance measurements (data not shown).

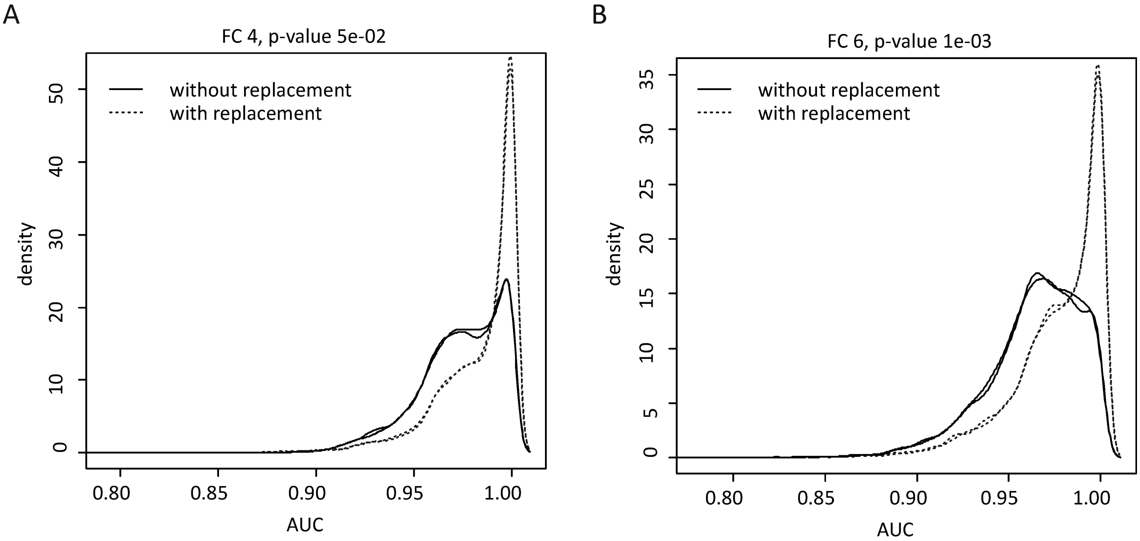


Figure 5.2.4.: Sample-split versus bootstrap predictions for small AML cohort Shown are AUC distributions for 10,000 iterations of TSA for different FC/p-value filter for sample split (sample drawing without replacement, black lines) and classical bootstrap (sample drawing with replacement, dashed lines), displayed as density plots.

5.2.2.4. Variation of trial simulation iterations

Overall, there are $\binom{150}{50} > 10^{40}$ ways to choose 50 elements (TS) from a set of 150 elements (AML PTS). As TSA is a rather computational intensive procedure, we compared results from 10,000 iterations to those obtained in 1000 and 100 permutations. When comparing the prediction performance, no significant differences were found between AUC distributions of different numbers of TSA permutations (Fig. 5.2.5A), the same is true for the other performance measurements (data not shown). Regarding feature rankings, 100 permutations resulted in higher variance, whereas the ranking obtained in 1000 permutations are highly comparable to the feature ranking from 10,000 iterations. Thus, in the following sections, 1000 iterations of TSA were used. In this way, TSA results for 150 samples can be obtained on a 64-bit Windows system with 8GB of RAM within few hours.

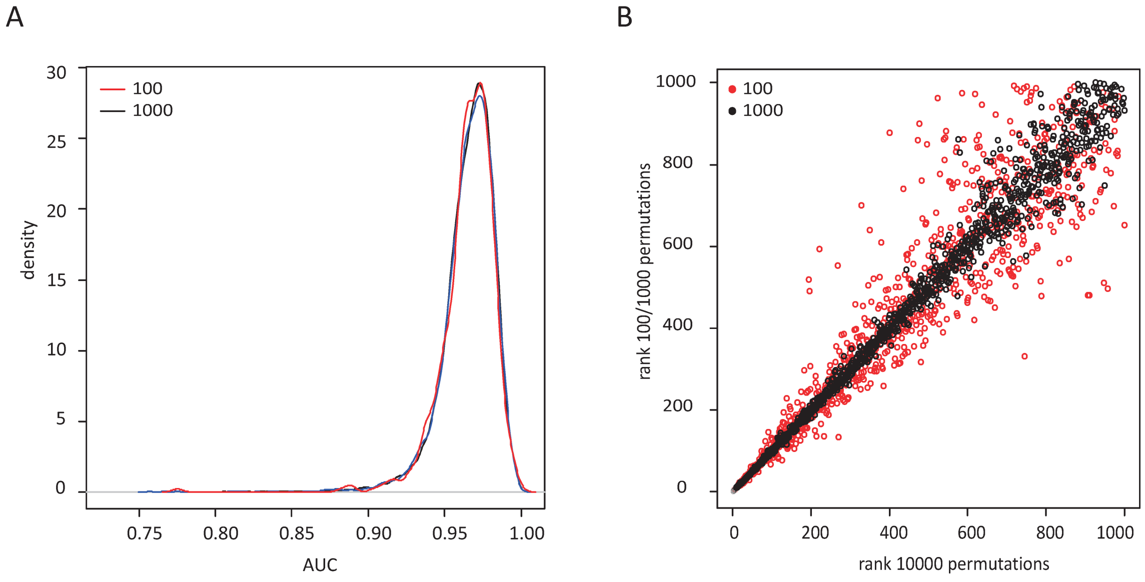


Figure 5.2.5.: Comparison of different TSA iterations
(A) Shown are AUC distributions of 10,000 (black), 1000 (blue) and 100 (red) permutations of TSA in the small AML cohort. (B) Feature ranking comparison: Shown is the feature ranking in 10,000 permutations from the 1000 highest ranked features obtained in 100 (red) and 1000 (black) permutations of TSA in the small AML cohort.

5.2.2.5. Comparison to the large data set

In clinical biomarker development results from small pilot trials are supposed to form the basis for larger validation trials, however, prediction of classifier performance in the larger cohorts is still an unsolved issue. Furthermore, classifier performance in larger cohorts is expected to improve (*Simon et al.*, 2003; *Ein-Dor et al.*, 2006). To capture the overall improvement by enlarging the cohorts (TS, V1, V2), we repeated the trial simulation approach on the complete AML data set (n=2013, Tab. 5.1). Again, linear SVM classification outperformed SVM performance with a radial kernel, LDA and PAM algorithms (Fig. 5.2.6A) as also observed in the small AML cohort. Although there was still a slight improvement of the spectrum of classifiers when increasing the feature size by changing feature selection cut-offs (Fig. 5.2.6B), all 6000 tests performed at least with an AUC of 0.977. Similar improvements were observed when reading out MCC, specificity or sensitivity performance (data not shown). In comparison to the small cohort, no NAs, indicating that no features passed the respective test cut-offs, were observed.

When directly comparing the initial PTS of the small AML data set with the complete AML data set it became clear that only the larger data set results in a sufficiently high AUC in the majority of classifiers developed. As shown in Fig. 5.2.7A, in the large AML cohort with 63.1% of all tests reached an AUC>0.99, 99.3% >0.98

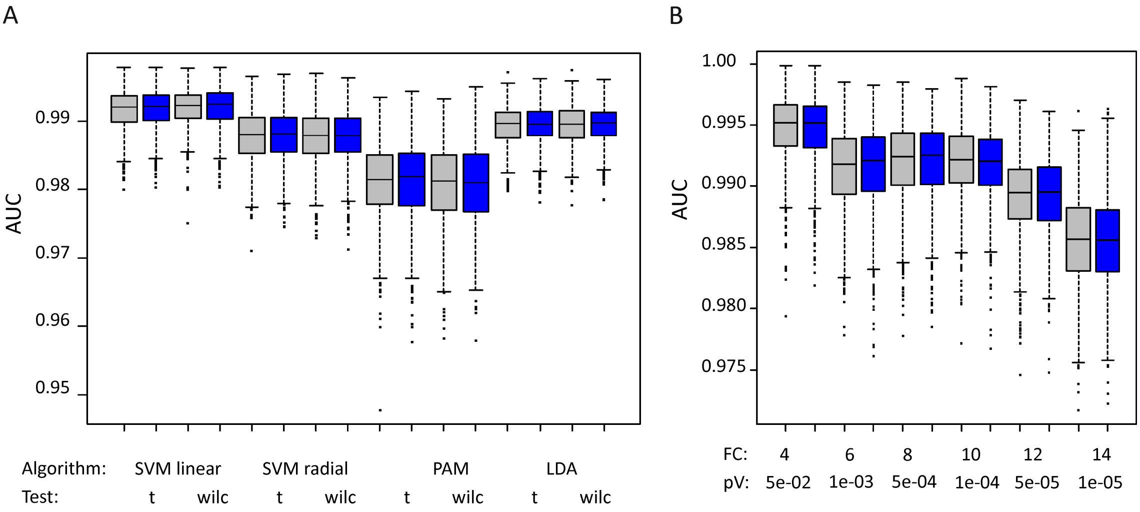


Figure 5.2.6.: Classification performance of large AML data set
AUC distributions shown as boxplots summarizing 1000 iterations of classification for (A) different classifier and feature selection algorithms (SVM with linear and radial kernel, PAM and LDA) in combination with a t-test or Wilcoxon test and (B) different feature size cut-offs for fold-change (FC)/p-value (pV) filter. The boundary of the box closest to zero indicates the 25th percentile, the line within the box marks the median and the boundary of the box farthest from zero indicates the 75th percentile. Whiskers above and below the box indicate the 90th and 10th percentiles. Outliers are plotted as dots.

and all classifiers resulted in an AUC>0.977. In fact, 81.2% of all classifiers generated in the small PTS showed an AUC>0.95 and 50% reached AUCs>0.967. Only 2.4% reached an AUC>0.99. Assessing the MCC showed similar results, while basically all classifiers generated in the large dataset reached an MCC >0.9, not even 60% of all classifiers within the small AML dataset reached an MCC>0.9 (Fig. 5.2.7B).

To elucidate whether the improvement in the larger dataset is associated with differences in feature distribution all transcripts (n=22,283) measured on the Affymetrix HG-U133A were evaluated for being part of at least one of 6000 classifiers (Fig. 5.2.7C). While a total of 2503 transcripts were part of at least one classifier in the PTS, only 606 transcripts were identified in the large data set. Even more striking, while no transcript was identified to be present in all classifiers in the small AML data set, 8 transcripts were present in all 6000 classifier in the large data set. When assessing feature size, an enormous variance became apparent in the small AML data set, while there was clearly less variance in the larger data set (Fig. 5.2.7D). Interestingly, reduced variance in feature size in the large data set was seen irrespective of filter criteria settings that determine the potential feature size. Together, these results support a robust gene expression profiling-based classifier as a test for primary diagnosis of AML with a sensitivity and specificity of greater 99.5%. At the same time these results also indicate that even initial PTS would require larger patient cohorts for robust classifier development, a requirement that can rarely be met.

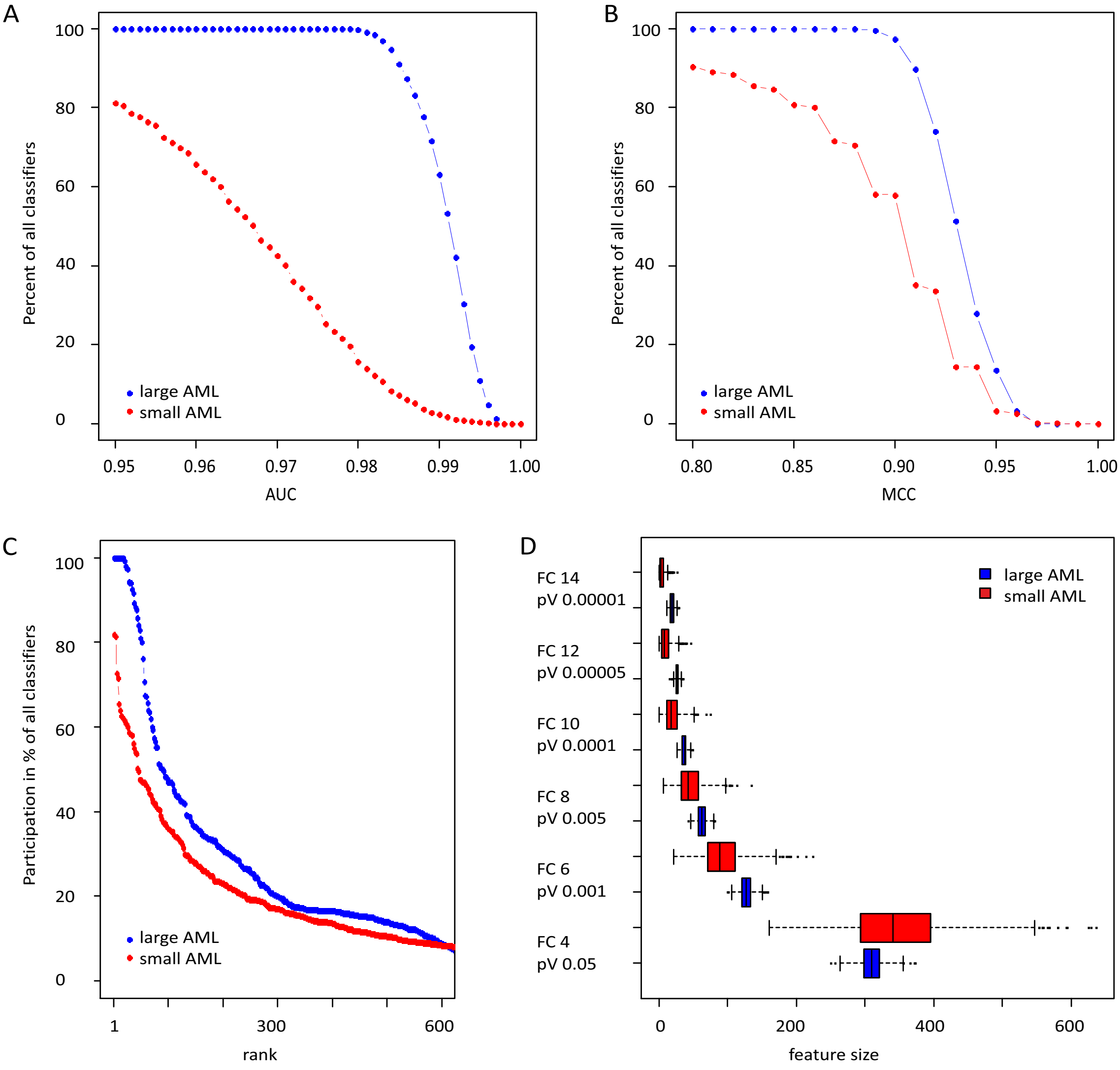


Figure 5.2.7.: Classifier performance and feature comparison between small and large AML cohort
(A+B) Shown is the percentage of all 6000 classifiers established in TSA reaching specific AUC (A) and MCC (B) thresholds in the small (red) and large (blue) AML cohort. (C) Shown is the percentage of feature participation in all 6000 classifier established in TSA. (D) Variance of feature size identified in 1000 TSA iterations summarized in boxplots. Both plots visualize the small AML chohort in red and the large AML cohort in blue.

5.2.2.6. Interpretation and validation of the final AML signature

We applied the feature ranking by ranks of the SVM weights to obtain a final AML signature from the large AML cohort (n=2013) as described in sec. 5.2.2.2. A complete annotation of the 50 highest ranked features from the large AML cohort is provided in Tab. A.6.

In order to validate this AML signature, we established another AML data set performed on the Affymetrix HG-U133 2.0 microarray. This study included a total of 2088 samples from 15 different studies not used in the previous analysis (Tab. 5.3). 1093 AML and 995 control samples were analyzed. Most control samples were healthy controls or unrelated diseases. As this data set was performed on a different platform, we only used transcripts also present on the HG-U133A platform (n=22,283 transcripts) and normalized both HG-U11A and HG-U133 2.0 arrays together using quantile normalization for batch-effect removal. A classifier was build on the HG-U133A data of 2013 samples using the weighted SVM AML signature and then the classifier was applied applied to the HG-U133 2.0 data set of 2088 samples.

Positives	Negatives	Further sample description	Reference
404		AML	*Mills et al.* (2009)
	302	Myelodysplastic syndrome	
96		AML	*Miesner et al.* (2010)
460		AML	*de Jonge et al.* (2010)
98		pediatric AML	
35		AML	*Silva et al.* (2009)
	9	liver transplant	*Martinez-Llordella et al.* (2007)
	8	healthy controls	
	6	Alzheimer/ Dementia	GSE18309
	3	healthy controls	
	39	Malaria vaccine trial	*Vahey et al.* (2010)
	240	Multiple sclerosis	*De Jager et al.* (2009)
	136	Juvenile idiopathic arthritis	*Barnes et al.* (2009)
	59	healthy controls	
	40	healthy controls	*Radom-Aizik et al.* (2009)
	27	Juvenile idiopathic arthritis	*Frank et al.* (2009)
	15	healthy controls	
	5	Steven-Johnson syndrome	*Chung et al.* (2008)
	59	Peridontitis	*Papapanou et al.* (2007)
	17	Juventile idiopathic arthritis	*Fall et al.* (2007)
	30	healthy controls	
1093	**995**	$\sum$	

Table 5.3.: HG-U133 2.0 AML data set
Summary of GEX studies included in the HG-U133 2.0 AML data set. Numbers shown are the actual number of samples used in the final data set. In total, this data set includes 2088 samples from 15 different studies.

For the 10, 20, 30, 50, 75, 100, 150 and 200 highest ranked features, a mean AUC performance of 0.9845 was reached (Tab. A.4), with a minimum AUC of 0.9671 from the 10 highest ranked features and a maximum AUC of 0,9972 from the 200 highest ranked features. Sensitivity ranged between 0.9469 and 0.9927 while specificity reached values between 0.8995 and 0.9899 at the threshold from the maximum Youden Index. Results were significantly higher in comparison to performance obtained when using the small AML cohort (Tab. A.5).

Furthermore, the best-performing 10, 20, 30 and 50 features performed also significantly better compared to 1000 random feature lists of equal size (Fig. 5.2.8), whereas for 75 and more features, results were not significant in comparison to 1000 random feature lists. This indicates that a smaller feature set is more informative and specified for AML prediction than a larger number of features, although all sets performed similarly well on the independent HG-U133 2.0 AML data set.

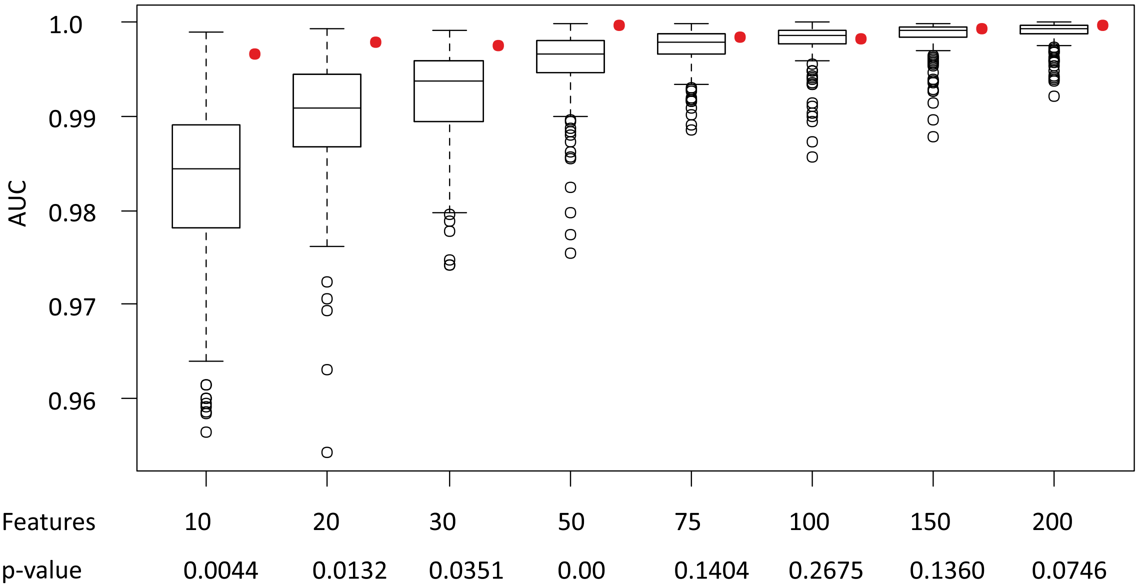


Figure 5.2.8.: Classifier performance on HG-U133 2.0 AML data set
Shown are box plots comprising 1000 AUCs obtained by using random feature lists of equal size for 10, 20, 30, 50, 75, 100, 150 and 200 features to build a classifier using the HG-U133A data applied to the HG-U133 2.0 data set. The real AUC using the specific SVM classifier from the weighted SVM ranking is depicted with red circles. The corresponding p-values are displayed below.

Next, the functional relevance of the obtained signature was analyzed. Gene Ontology (*Ashburner et al.*, 2000) pathway analysis by GOrilla (*Eden et al.*, 2009) revealed a significant enrichment for immune system processes as well as related pathways such as T-cell, lymphocyte and leukocyte activation (see Tab. 5.4). Furthermore, when performing enrichment analysis of Disease Ontology annotation (*Osborne et al.*, 2009) associated with the 200 highest ranked AML features using the FunDO tool (*Du et al.*, 2009), leukemia was the most relevant significantly

GO term	Description	P-value
GO:0002376	immune system process	6.52E-21
GO:0006955	immune response	1.46E-12
GO:0002682	regulation of immune system process	9.7E-12
GO:0051249	regulation of lymphocyte activation	2.45E-10
GO:0050863	regulation of T cell activation	2.79E-10
GO:0050870	positive regulation of T cell activation	9.1E-10
GO:0002696	positive regulation of leukocyte activation	9.19E-10
GO:0050867	positive regulation of cell activation	1.66E-9
GO:0051251	positive regulation of lymphocyte activation	2.01E-9
GO:0050896	response to stimulus	2.09E-9

Table 5.4.: GO enrichment analysis of 200 top ranked AML features. Shown are the 10 most significantly enriched GO terms.

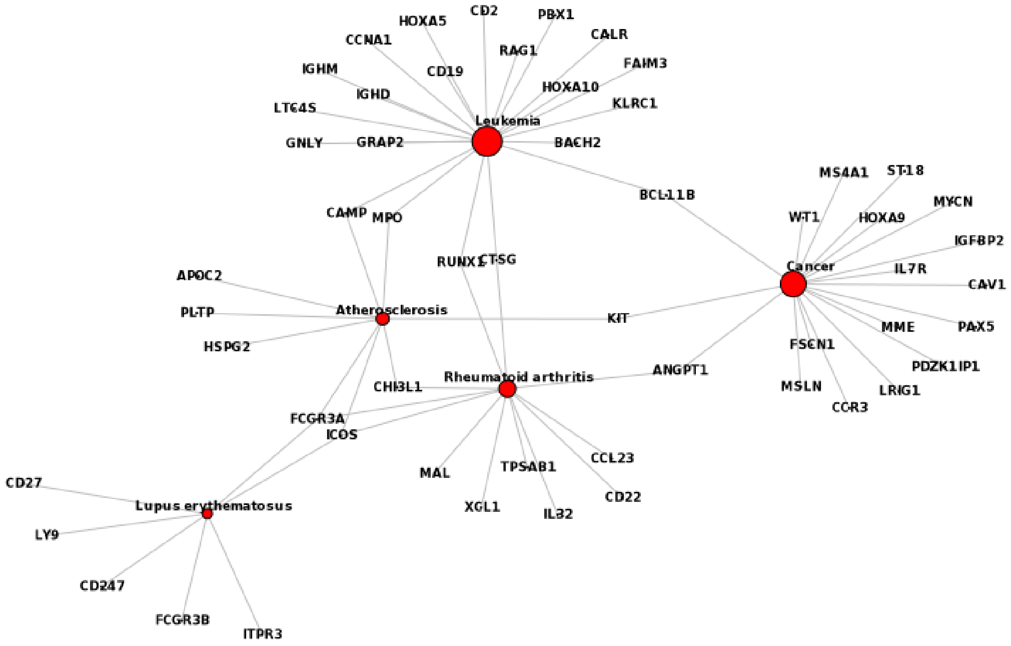
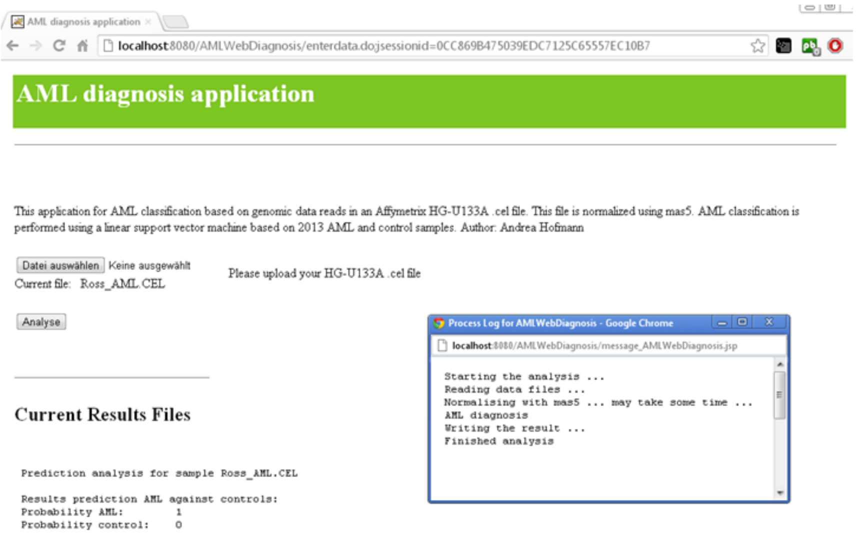


Figure 5.2.9.: Disease Ontology enrichment analysis of AML signature
Disease Ontology enrichment analysis for the 200 highest ranked AML features. Shown are the top five diseases and the genes in the list which map to them. The sizes of the disease nodes are proportional to the number of edges.

enriched disease term (Bonferroni corrected p-value $< 10^{-18}$), followed by cancer (Bonferroni corrected p-value $< 10^{-8}$) as visualized in Fig. 5.2.9. These results indicate a functional relevance of the obtained AML signature and a phenotype-related feature ranking.

5.2.2.7. Implementation of a web-based tool for AML classification

To make the established AML data set and the validated signature available for scientists who are not familiar with a command line based environment like the R language, I implemented a web-based AML diagnostic system prototype using RWui (*Newton et al.*, 2011). Within this application AML primary diagnosis is available and classification prediction probabilities for a new sample are reported (Fig. 5.2.10). An Affymetrix microarray HG-U133A cel-file has to be uploaded and is normalized within the system. Then classification analysis starts using a linear SVM established in the total AML HG-U133A data set (n=2013 samples) with the highest 200 features from sec. 5.2.2.6. During the workflow the diagnostic software provides the user process information and the results can be displayed directly or downloaded. In a clinical application, the diagnostic software would ask the user to fill in detailed sample information and clinical covariables e.g. molecular characteristics and FAB status. The system could also offer AML subclassification or chronic leukemia differential diagnosis if sample size and therefore statistical power will be sufficient.


AML diagnosis application
localhost:8080/AMLWebDiagnosis/enterdata.do;jsessionid=0CC869B475039EDC7125C65557EC10B7

AML diagnosis application

This application for AML classification based on genomic data reads in an Affymetrix HG-U133A .cel file. This file is normalized using mas5. AML classification is performed using a linear support vector machine based on 2013 AML and control samples. Author: Andrea Hofmann

Datei auswählen Keine ausgewählt
Current file: Ross_AML.CEL

Please upload your HG-U133A .cel file

Analyse

Current Results Files

```
Prediction analysis for sample Ross_AML.CEL

Results prediction AML against controls:
Probability AML:        1
Probability control:    0
```

Process Log for AMLWebDiagnosis - Google Chrome

localhost:8080/AMLWebDiagnosis/message_AMLWebDiagnosis.jsp

```
Starting the analysis ...
Reading data files ...
Normalising with mas5 ... may take some time ...
AML diagnosis
Writing the result ...
Finished analysis
```

Figure 5.2.10.: AML diagnosis tool

5.2.2.8. Further research on AML subclassification

In principle, genomic profiling using microarray technology could be used for primary diagnosis, differentially diagnosis, subclassification of diseases, therapy outcome, prognosis and prediction of disease. In the previous section, we proved that primary AML diagnosis is possible with high accuracy. Currently subclassification of patients with AML is done using a range of molecular markers. Hence the presence or absence of recurrent cytogenetic aberrations is used to identify the appropriate therapy. Several studies already reported the successful establishment of transcriptional-based classifier for disease subclassification (*Valk et al.*, 2004; *Ross et al.*, 2004; *Bacher et al.*, 2009a,b). We assessed subclassification based on cytogenetic aberrations (a summary of subgroups analyzed is provided in Tab. A.1) by TSA. Due to small sample sizes, we only draw one TS and one VS from the entire data in the ratio 1:1. As shown in Fig. 5.2.11, prediction performance depends on the subtype analyzed in this meta-analysis. While favorable risk groups such as t(15;17), t(8;21) and inv(16) result in mean AUC values > 0.8 the chromosomal abnormality 11q23 assigned as an unfavorable risk group is more difficult to predict indicating that either the current classification system does not entirely reflect the molecular heterogeneity of the disease or sample sizes are too small.

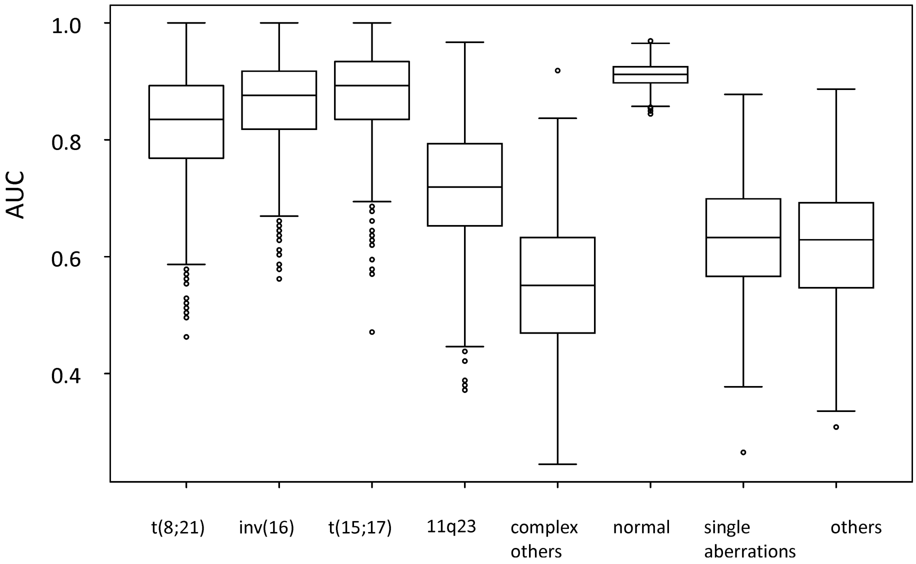


Figure 5.2.11.: TSA for AML subclassification
Shown are the AUC values of 1000 TSA permutations for 8 different subgroups defined by cytogenetic aberrations. Due to small sample sizes, only one training and one validation set were drawn from the entire data at a ratio 1:1.

5.2.3. Adaptive learning approach allows classifier optimization and further sample size estimation

The MAQC-II study and others (*Ein-Dor et al.*, 2005; *Shi et al.*, 2010) have clearly established that many classifiers with similar performance can be developed from a given data set. Therefore, static use of the initial best-performing signatures might not be the best choice in the end, once clinical application is considered. Especially when using high-throughput gene expression profiling, completely specified and analytically validated biomarker candidates may not be feasible before its use in pivotal clinical trials. Adaptive trial designs for phase III trials when no classifier is available at the start of the trial have been proposed (*Freidlin and Simon*, 2005; *Jiang et al.*, 2007; *Simon*, 2008). Consequently, approaches that lead to continuous optimization of existing classifiers should be evaluated irrespective of current regulatory issues concerning such strategies (*FDA*, 2005). In essence, continuous classifier optimization can be achieved by adaptive learning techniques opening completely novel opportunities when integrating large parallel biological data for medical diagnostics.

We modeled and tested our data-driven adaptive learning approach for diagnostic classifier development as follows (see Fig. 5.2.12): First, we divided our compiled large AML data set (n=2013) into one set for adaptive learning (ALS) and two external independent VSs, each cohort equally comprising n=671 samples. The ALS was used for adaptive learning assessment. An initial TS of 50 samples (25 AML, 25 controls) was randomly drawn from the ALS. Internal classifier validation was performed in this set by TSA (100 iterations) as described in sec. 5.2.2. The mean AUC for 100 TSA permutations was saved. A classifier build in the entire initial TS was applied to the two external VSs (external validation). In the next step, further 20 samples randomly taken from the remaining samples in the ALS were added to the initial TS and internal as well as external validation was performed using this set of 70 samples, resulting in a mean AUC for internal and two AUCs for external validation. Next, 20 new samples were included into a 90-sample comprising set, internal and external validation was applied and this procedure was repeated for sequentially adding multiple of 20 samples until all ALS samples were used (32 sequential draws). To ensure results were not depending upon a specific sample sequential arrangement, adaptive learning with the fixed ALS and VSs was repeated 100 times for bias reduction. Thus in summary, 100 mean AUCs for internal and 2 AUCs for external validation were saved for each sample size.

As expected, for smaller numbers of samples the AUC was lowest, but increased steadily with increasing numbers of samples (visualized by the boxplots of Fig. 5.2.13). As underscored by higher AUC values in the two external VSs for each given sample size, TSA underestimated the statistical performance of the developed classifiers (brown and orange lines in Fig. 5.2.13).

Another aspect in gene signature development that is still controversial is the sample size of initial pilot studies (*Ein-Dor et al.*, 2006). While prediction performance of the clinical dataset within MAQC-II were judged to be more difficult endpoints, it was not formally addressed whether the sample size of the data sets was too small to reach higher statistical power given the underlying biological difference between the classes within the data sets. Adaptive learning can be used for curve sketching allowing the calculation of the number of patients required to reach a certain level of statistical performance.

Hence next, we evaluated the adaptive learning results to determine minimal sample

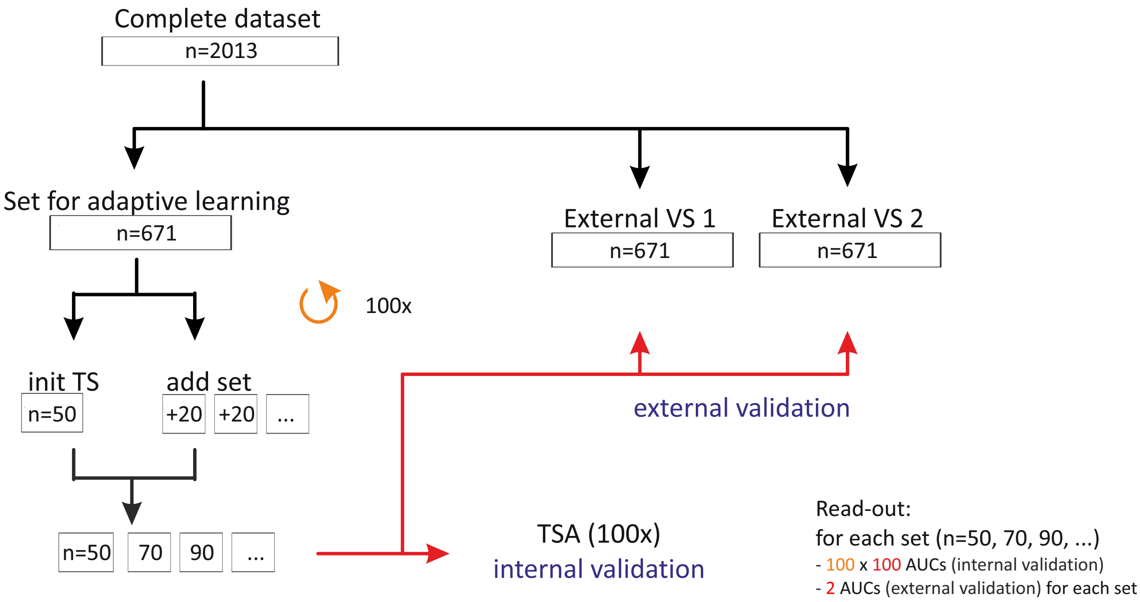


Figure 5.2.12.: Schematic view of the adaptive learning design
The large AML data set is divided into one set for adaptive learning (ALS) and two external independent VSs, each cohort equally comprising n=671 samples. An initial TS of 50 samples (25 AML, 25 controls) is randomly drawn from the ALS. Internal classifier validation is performed in this set by TSA (100 iterations). A classifier build in the entire initial TS is applied to the two external VSs (external validation). Next, further 20 samples randomly taken from the remaining samples in the ALS are added to the initial TS and internal as well as external validation was performed using this set of 70 samples. This procedure is repeated by sequentially adding multiple of 20 samples until all ALS samples are used. Adaptive learning with the fixed ALS and VSs was repeated 100 times for bias reduction. Thus in summary, 100 x100 AUCs for internal and 2 AUCs for external validation are read out for each sample size.

size required for validation cohorts. Therefore, we fitted a cubic smooth spline curve to the summarized mean-of-the-mean AUCs (see sec. 3.2.7), indicated by the blue line of Fig. 5.2.13. The smoothing spline estimate $\hat{f}$ is used to predict the maximum value x_{max} to calculate the maximum performance of a classification model for a specific endpoint. Sample sizes for pivotal validation cohorts were estimated at 90% and 95% of this maximum by calculating $\hat{f}(0.9 * x_{max})$ and $\hat{f}(0.95 * x_{max})$, respectively (red lines in Fig. 5.2.13). In this AML model, a trial with 290 patients would have been estimated by adaptive learning to reach 90% of maximum statistical performance, while 390 new patients would have been required to reach 95%.

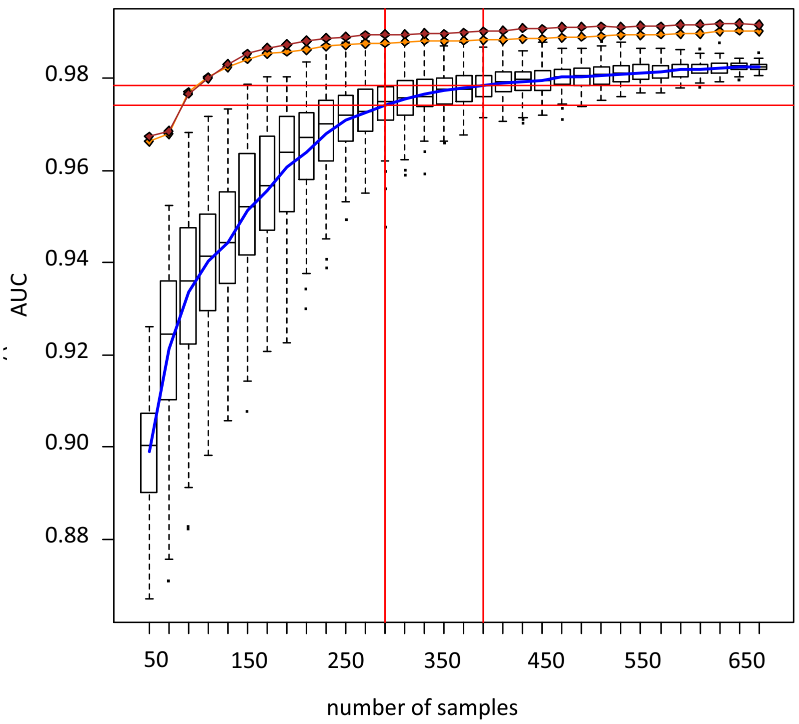


Figure 5.2.13.: Adaptive learning for continuous classifier optimization and sample size estimation
Prediction performance obtained during adaptive learning is summarized for internal and external validation. Shown are boxplots from 100 TSA for each adaptive learning set (internal validation) and external validation estimates for external VS1 (brown) and VS2 (orange). A blue line displays the smooth cubic spline fit to the internal validation AUCs. 90% and 95% sample size estimations are marked with red lines.

5.2.4. Adaptive learning and trial simulation of extrapolated data

In initial pilot trials, sample sizes are usually too small to establish completely specified and analytically validated biomarker candidates (*Ein-Dor et al.*, 2006). The generation of more samples often is not possible or expensive and time-consuming. With this practical limitation in mind, an alternative might be the inclusion of artificial samples, referred to as data oversampling or extrapolation. The idea is to simulate new samples according to the distribution of given samples. We choose a rather simple feature-by-feature approach using metric relationships from two randomly chosen samples to generate new samples (see sec. 3.2.6 for details). Here, this oversampling was used for sample size estimation of larger pivotal validation studies based on an initial small data cohort, which would be beneficial to accelerate clinical translation of diagnostic GEX studies and parameters were chosen to increase the variance within the data set, this is to increase the range of expression values for expressed transcripts and to decrease homogeneity in the simulated data set.

Based on the PTS of AML data (n=150), we generated a total of 370 AML and 370 controls, resulting in a new data set with a total of 890 samples including the original 150 real samples. When applying adaptive learning to the extrapolated data set, we observed a higher variance of AUC, particularly at lower sample size. Moreover, the means of mean AUCs was lower when compared to the real data set. In addition, patient estimation for 90% respectively 95% statistical performance revealed higher numbers (450 and 630 patients) than observed within the real data set (290 and 390 patients, Fig. 5.2.14).

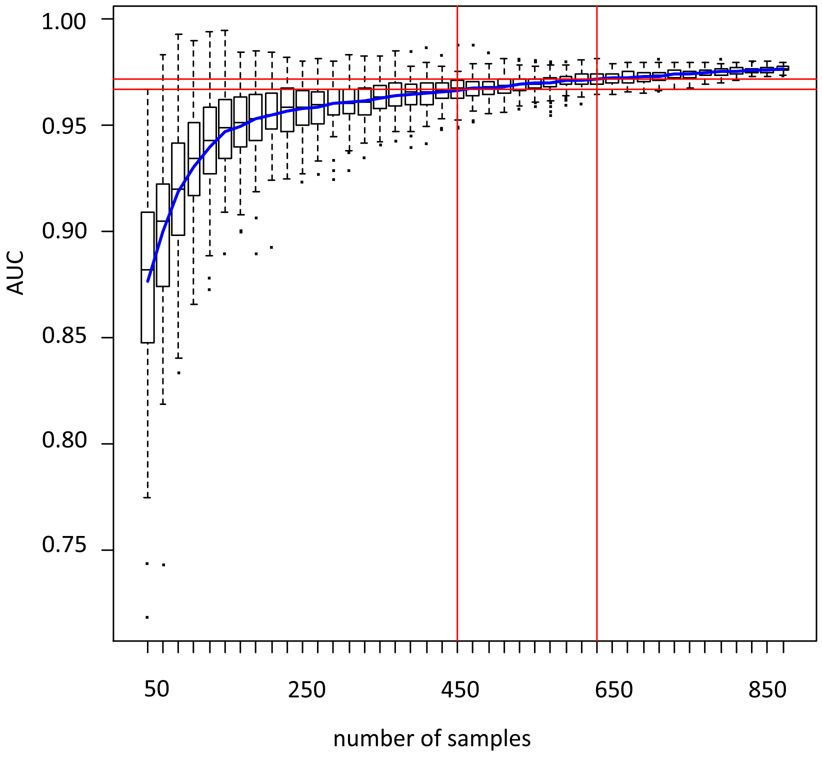


Figure 5.2.14.: Adaptive learning of extrapolated AML data set
Shown are boxplots from 100 TSA for each adaptive learning set. A blue line displays the smooth cubic spline fit to the mean-of-the-mean AUCs. 90% and 95% sample size estimations are marked with red lines.

For completion, we also applied a more sophisticated approach for microarray data extrapolation proposed by *Parrish et al.* (2009). This method is based mathematical transformations of the underlying expression measures such that the transformed variables follow approximately a Gaussian distribution to estimate associated parameters. Using this model, we generated 717 AML and 1296 control samples and applied our adaptive learning approach. As visualized in Fig. B.0.5, increasing sample sizes did not lead to higher AUC values. Remarkable, the prediction performance remains similar independent of an underlying cohort size. Apparently, extrapolated samples were already more similar than the real underlying data set making this approach not applicable for sample simulation in adaptive learning based on a small pilot trial.

Overall, although our data extrapolation approach overrates the sample sizes for larger pivotal validation cohorts, it is well suitable to estimate patient numbers based on the distribution of a given initial data set required for biomarker approval. Furthermore, while we have identified more than 4,000 publically available sample data suitable to develop a very reliable biomarker for primary molecular diagnosis of AML, this analysis clearly demonstrated that less than 80% of patient samples already would have been sufficient to present this test to the drug authorities.

We also performed TSA on the extrapolated AML data set to compare these results to the small and large AML prediction performance. Applying 1000 iterations of TSA to the extrapolated data set revealed a substantially higher prediction performance compared to the small data set (Fig. 5.2.1D), however, performance of the large data set (Fig. 5.2.6B) was not reached, as shown by the percentage of classifiers with high AUCs (Fig. 5.2.15A). The number of features being part of up to 6000 classifiers increased similarly to the large data set (Fig. 5.2.15B). More importantly, when comparing the weights SVM ranking of the features obtained in TSA, a higher concordance between the large and extrapolated cohort could be shown (Fig. 5.2.15C). In this regard the extrapolated and the large data set were more similar than the large and the small data set. Next, we investigated the classification ability of these features and data set. Therefore, we established classifiers using the highest ranked features (best ranked 10, 20, 30, 50, 75, 100, 150 and 200 features; n=8 cut-offs in total) derived from extrapolation with both the original small PTS and the new extrapolated data set and applied the classifier to the remaining samples of the large AML data. Compared to the AUCs obtained by the original PTS, neither highest ranked extrapolated features on the original PTS nor highest ranked extrapolated features with the extrapolated data set reached comparable prediction performance.

Thus, despite higher similarity between extrapolated and large cohort in terms of feature stability, feature ranking and prediction performance within each data set, this concordance did not enhance prediction performance on a data set not used in the analysis.

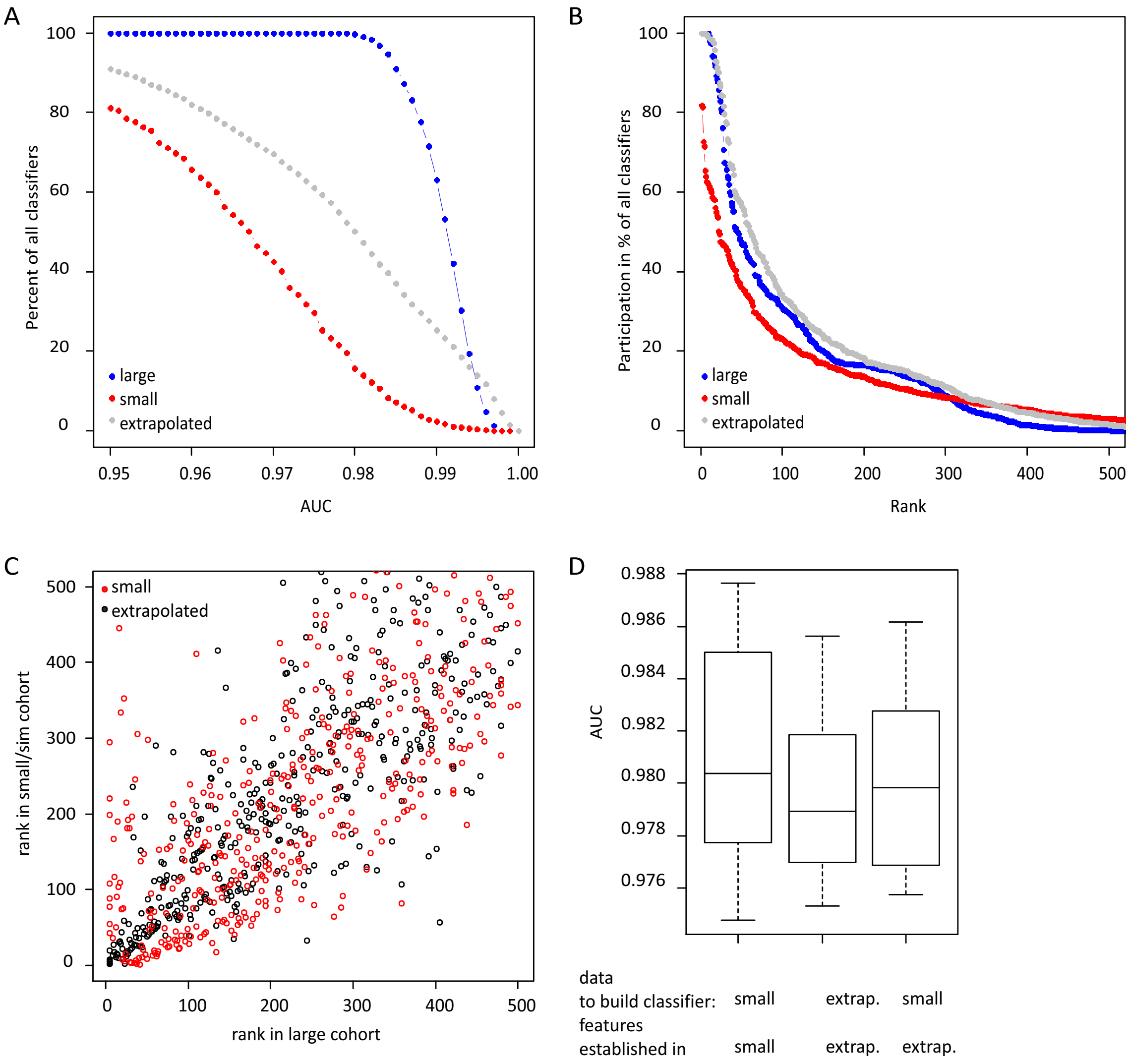


Figure 5.2.15.: Performance of extrapolated data set

(A) Percentage of all 6000 classifier established in TSA reaching specific AUC thresholds in the small (red), extrapolated (grey) and large (blue) data. (B) Percentage of feature participation in all 6000 classifier established in TSA in the small (red), extrapolated (grey) and large (blue) data. (C) Feature ranking of large data versus feature ranking in the small data (red) and extrapolated data (black) showing the top 500 ranked features. (D) AUC prediction results using the best ranked 10, 20, 30, 50, 75, 100, 150 and 200 features (n=8) from the small and extrapolated data as a classifier build in the small and extrapolated (sim) data set applied to the remaining samples of the large AML cohort. All 8 AUC values are summarized in box plots for 3 different combinations of data used to build the classifier and data used to get the feature ranking.

5.2.5. Performance and predictions of other disease data sets

We next extended our findings by applying the established methods TSA and adaptive learning to three additional data sets of blood-based gene expression profiling. All sets were generated on a different platform, the Illumina bead-based array system. Identification of patients with early stage lung cancer (NSCLC), active TB, or HIV infection were chosen as endpoints. A summary of all sets used is given in Tab. 5.5. For each setting feature selection was varied by different FC/p-value filter cut-offs (n=6) and TSA applied thereby receiving a total of 6.000 different cohort simulations per data set.

Data set	Platform	Number of samples	Positives	Negatives	Reference
complete AML	Affymetrix HG-U133A	2013	717	1296[a]	Tab. 5.1
PTS AML	Affymetrix HG-U133A	150	75	75	random samples from complete AML data
val. AML	Affymetrix HG-U1332.0	2088	1083	9956[a]	Tab. 5.3
NSCLC	Illumina WG6 V2	233	95	138[b]	*Zander et al.* (2011)
HIV	Illumina WG6 V2	257	106[c]	151[d]	Debey-Pascher, Tab. A.2
TB	Illumina HT12 V3	147	54	93[e]	*Berry et al.* (2010)

Table 5.5.: Microarray data sets used for model development and validation Summary of GEX data sets are given. [a]Other leukemia, other diseases, healthy controls, [b]Hospital-based controls, healthy controls, [c]HIV, LTNP, [d]Other diseases, healthy controls, [e]Latent TB, healthy controls.

5.2.5.1. NSCLC

First, we used the NSCLC data set used to establish our preferred classification workflow in chapter 4. Similar to AML, TSA revealed a decrease in performance with more stringent FC/p-value filter criteria, still overall mean AUC values of 0.8776 were reached (Fig. 5.2.16A), comparable to the AUCs of 0.824 and 0.977 obtained in chapter 4 for VS1 and VS2. Few features passed the FC$\geq$ 3/p-value<0.0001 filter cut-off (21%). Mean sensitivity $\geq$ 0.84 and mean specificity $\geq$ 0.77 could be obtained at the maximum Youden Index (YI). Next, we extrapolated the original NSCLC data set to 1160 samples and applied our data-driven adaptive learning approach by applying TSA as internal validation to each of a total of 30 steps. Sample size estimations by cubic spline fitting for a pivotal validation cohort were 610 individuals to reach 90% of highest performance and further 730 samples would be required to get 95% performance (Fig. 5.2.16B).

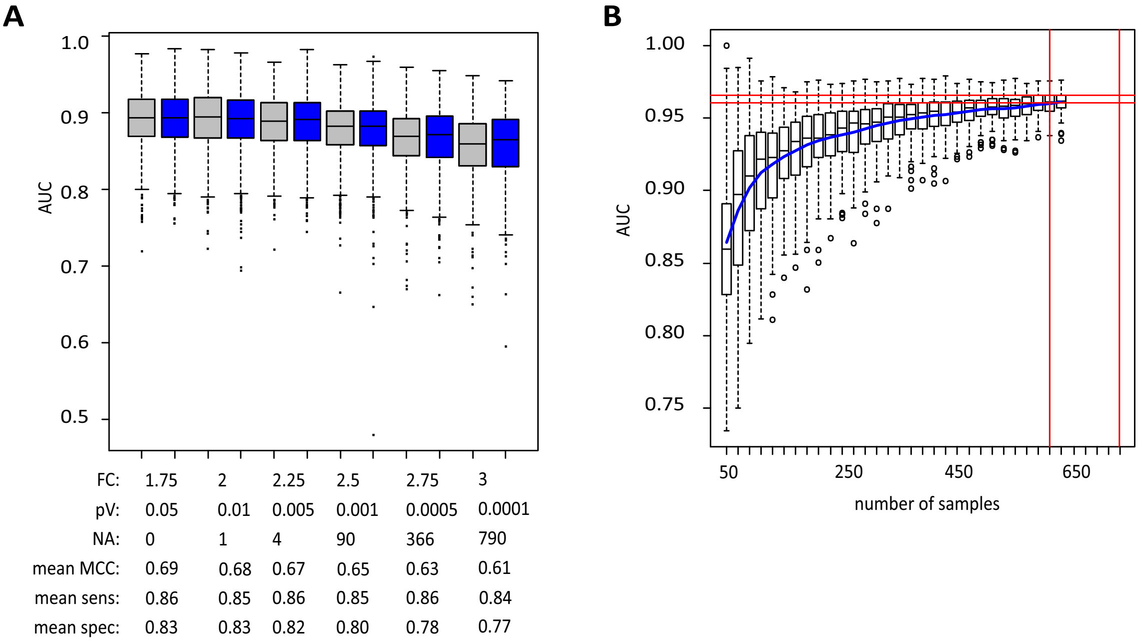


Figure 5.2.16.: TSA and adaptive learning of NSCLC data set
A: 1000 TSA permutations for 6 different FC/p-value feature selection cut-off combinations. Shown are the boxplots summarizing AUC prediction performance. NA values indicate the number of tests without features passing the filter cut-offs. Mean MCC, sens and spec for both cohorts are given below. B: Adaptive learning using simulated data based on the original NSCLC cohort. Boxplot comprise 100 mean-of-the-mean AUCs of 100 TSA for each adaptive learning set. A blue line mark the fitted cubic spline curve, while red lines indicate 90% and 95% estimations.

Among the 200 highest ranked features derived from weighted SVM ranking in TSA (see sec. 5.2.2.2), 66 were enriched for disease-terms. GO pathway analysis reveals on one side an association with stimulus response but also an enrichment of immune and immune system processes, as shown in Tab. 5.6. Tab. A.7 provides a detailed annotation of the 50 highest ranked features. When comparing the top ranked features obtained by TSA to the 484-signature from chapter 4, there was only a 9% overlap to the top 500 ranked features.

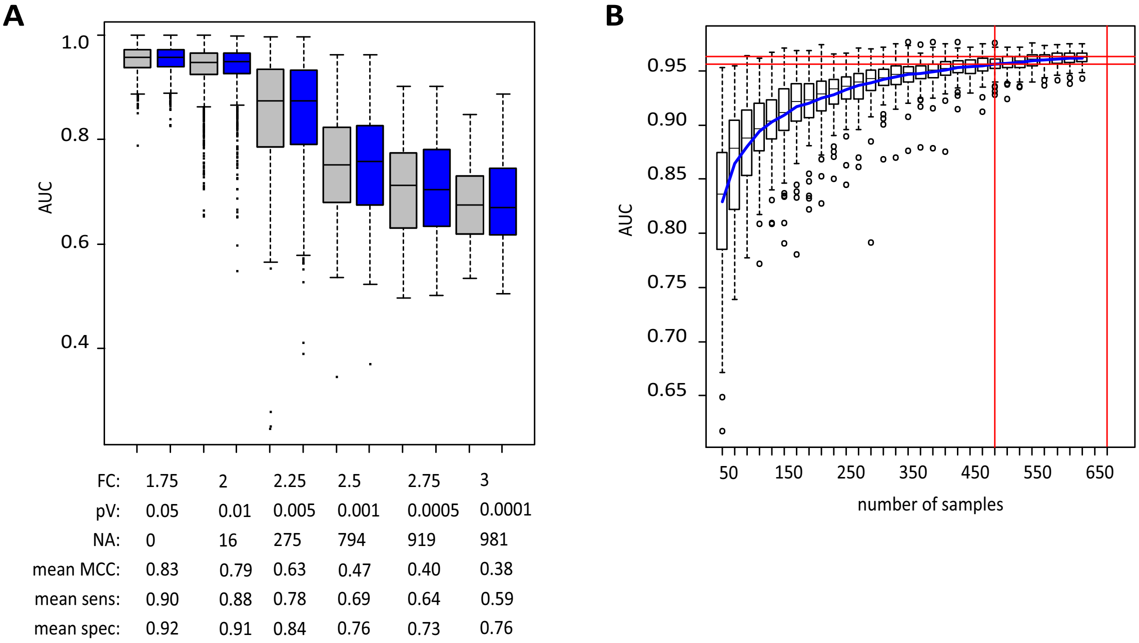
GO term	Description	P-value
GO:0009607	response to biotic stimulus	6.3E-12
GO:0051707	response to other organism	7.59E-12
GO:0035821	modification of morphology or physiology of other organism	2.75E-11
GO:0031640	killing of cells of other organism	1.16E-10
GO:0006955	immune response	2.08E-10
GO:0002376	immune system process	2.64E-10
GO:0006952	defense response	2.86E-10
GO:0001906	cell killing	2.25E-9
GO:0009620	response to fungus	2.25E-9
GO:0050832	multi-organism process	3.43E-9

Table 5.6.: GO enrichment analysis of 200 top ranked NSCLC features. Shown are the 10 most significantly enriched GO terms.

5.2.5.2. HIV

Another GEX data set was established in our group to identify a cohort including both HIV+ and LTNP patients in comparison to control samples (see Tab. A.2). After HIV infection, the individual genetic background plays a central role for the variable disease progression towards Acquired Immune Deficiency Syndrome (AIDS). LTNPs are rare represented individuals in the HIV infected patient population ($1-2\%$) having AIDS-free survival without antiretroviral therapy for more than 10-15 years (*Sheppard et al.*, 1993). In this project, we assessed primary HIV diagnosis and included the LTNP samples (n=27) in the HIV cohort (n=79) in order to improve HIV diagnosis in the clinic. Our control cohort comprises healthy controls (n=88) as well as patients with other inflammatory diseases (34 patients with sepsis, 29 patients with scleroderma).

Again, prediction performance was increased when increasing the feature size for classifier generation (Fig. 5.2.17A). Stringent filter selection cut-offs of FC$\geq$ 2.5/p-value<0.001 and above resulted in few features passing those filters (20.6%) and hence, less classifiers could be build. For the filters FC$\geq$ 2.75/p-value<0.005 and FC$\geq$ 3/p-value<0.0001 only 8.1% and 1.9% of all sample subsets could be used to build classifiers for the chosen filter criteria. For the feature filters FC$\geq$ 1.75/p-value<0.05 a mean sensitivity $\geq$ 0.90 and mean specificity $\geq$ 0.92 at the maximum YI were observed. As a next step we applied our adaptive learning approach to the simulated HIV data sets of 1280 samples to estimate the number of samples required for a validation and approval trial. We generated 300,000 classifier applying TSA in 30 steps. Based on curve sketching of adaptive learning prediction results we expect the development of a test to detect HIV with an AUC exceeding 0.95. To reach this goal, adaptive learning predicts the requirement for a validation trial with 490 individuals to obtain 90% of the maximum statistical performance and 670 individuals for 95% (Fig. 5.2.17B).



FC:	1.75	2	2.25	2.5	2.75	3
pV:	0.05	0.01	0.005	0.001	0.0005	0.0001
NA:	0	16	275	794	919	981
mean MCC:	0.83	0.79	0.63	0.47	0.40	0.38
mean sens:	0.90	0.88	0.78	0.69	0.64	0.59
mean spec:	0.92	0.91	0.84	0.76	0.73	0.76

Figure 5.2.17.: TSA and adaptive learning of HIV data set
A: 1000 TSA permutations for 6 different FC/p-value feature selection cut-off combinations. Shown are the boxplots summarizing AUC prediction performance. NA values indicate the number of tests without features passing the filter cut-offs. Mean MCC, sens and spec for both cohorts are given below. B: Adaptive learning using simulated data based on the original HIV cohort. Boxplot comprise 100 mean-of-the-mean AUCs of 100 TSA for each adaptive learning set. A blue line mark the fitted cubic spline curve, while red lines indicate 90% and 95% estimations.

When assessing the 200 highest ranked features in the HIV data set (see Tab. A.8 for detailed annotation of 50 top features) identified by using the weights of SVMs in TSA, 55 were found to be associated with diseases. Interestingly, enriched GO terms comprised multi-organism processes such as morphology or physiology modification of other organism, killing cells of other organism as well as response to biotic stimulus and bacterium. Furthermore, viral transcription and viral infectious cycle were included into significantly enriched GO terms of the HIV signature, thus supporting a biological meaningful function of the derived HIV signature as well (Tab. 5.7).

5.2.5.3. TB

We further analyzed a recently published data set that was used to establish a diagnostic signature for ATB in whole-blood (*Berry et al.*, 2010). TB is an infec-

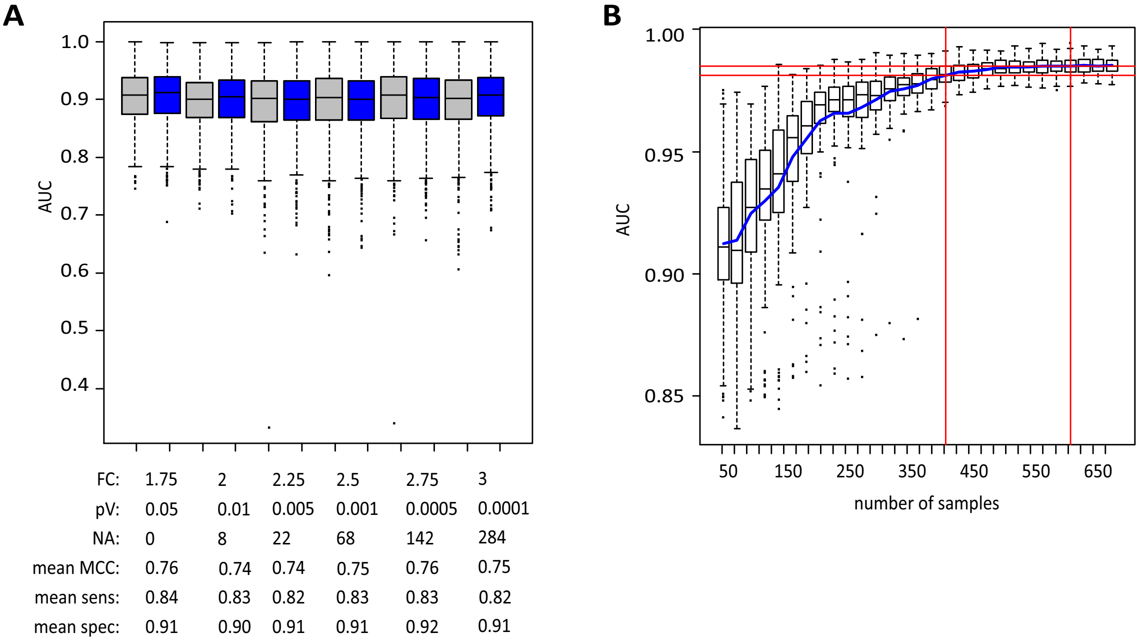
GO term	Description	P-value
GO:0035821	modification of morphology or physiology of other organism	5.24E-8
GO:0031640	killing of cells of other organism	9.87E-8
GO:0009607	response to biotic stimulus	6.91E-7
GO:0001906	cell killing	7.26E-7
GO:0019083	viral transcription	8.97E-7
GO:0006415	translational termination	1.16E-6
GO:0019058	viral infectious cycle	1.61E-6
GO:0009617	response to bacterium	2.18E-6
GO:0043624	cellular protein complex disassembly	2.56E-6
GO:0043241	protein complex disassembly	2.75E-6

Table 5.7.: GO enrichment analysis of 200 top ranked HIV features. Shown are the 10 most significantly enriched GO terms.

tious disease caused by *Mycobacterium tuberculosis* strains. Most infections remain asymptomatic, termed latent TB, and about $5-10\%$ progress to active disease, which can be fetal if left untreated (*Lawn and Zumla*, 2011). TB diagnosis is made by finding *M. tuberculosis* bacteria in a clinical sample and current tests cannot identify individuals with ATB (*Barry et al.*, 2009). We assessed primary ATB diagnosis by comparing ATB against latent TB (LTB) and control samples in order to establish whole-blood based transcriptional biomarkers to identify individuals who will develop the active disease (Tab. A.3).

In comparison to the TSA results from AML, NSCLC and HIV data, TB prediction performance was not dependent on feature selection cut-off. More features passed the stringend feature selection cut-offs of FC$\geq$ 3/p-value<0.0001 (71.8%). The statistical performance was nearly similar for all 6 cut-offs (Fig. 5.2.18A) with a mean AUC of 0.8974. In the original paper, a sensitivity of 61.67% and a specificity of 93.75% were reported (*Berry et al.*, 2010). In our setting, mean sensitivities ≥ 0.82 and mean specificities ≥ 0.90 were obtained at the maximum YI, showing a high increase in case detection rates by TSA. For application of our data-driven adaptive learning approach we extrapolated the original TB data set to a set of 879 samples to estimate the number of samples required for a validation and approval trial. In total, 290,000 classifier were build applying 100 TSA permutations in 29 steps. Sample size estimations were 410 individuals to reach 90% of the maximum statistical performance and 510 individuals for 95%. Maximum performance exceeded for a TB blood test was 0.97 (Fig. 5.2.17B).

Pathway analysis of the 200 most informative TB features revealed that 45 features were disease related genes. Functional GO analysis found an enrichment of GO terms related to immune system processes such as different immune responses, response to cytokine stimulus and B cell receptor signaling pathway (Tab. 5.8). A complete



FC:	1.75	2	2.25	2.5	2.75	3
pV:	0.05	0.01	0.005	0.001	0.0005	0.0001
NA:	0	8	22	68	142	284
mean MCC:	0.76	0.74	0.74	0.75	0.76	0.75
mean sens:	0.84	0.83	0.82	0.83	0.83	0.82
mean spec:	0.91	0.90	0.91	0.91	0.92	0.91

Figure 5.2.18.: TSA and adaptive learning of TB data set
A: 1000 TSA permutations for 6 different FC/p-value feature selection cut-off combinations. Shown are the boxplots summarizing AUC prediction performance. NA values indicate the number of tests without features passing the filter cut-offs. Mean MCC, sens and spec for both cohorts are given below. B: Adaptive learning using simulated data based on the original TB cohort. Boxplot comprise 100 mean-of-the-mean AUCs of 100 TSA for each adaptive learning set. A blue line mark the fitted cubic spline curve, while red lines indicate 90% and 95% estimations.

annotation of the 50 highest ranked features is given in Tab. A.9. In contrast to NSCLC, there was an overlap of 42% of the top 400 ranked TSA features with the recently published biology-driven 393-feature list (*Berry et al.*, 2010) strongly suggesting that the biology-driven approach to obtain a classifier revealed high accuracy for the most important features for classifier generation.

5.3. Summary and Discussion

In summary, we established a comprehensive workflow to accelerate the clinical translation of gene expression based biomarker. Overall, 1,800,000 models were evaluated (Tab. A.10). First, we developed a randomized clinical trial simulation approach (TSA) to better judge the variation of classification performance one can obtain from a small pilot trial. Most informative features for classifier generation can be identified using ranksums of all signatures of TSA further weighted by underlying SVMs. This approach was developed based on a small AML pilot trial randomly

GO term	Description	P-value
GO:0002376	immune system process	4.68E-9
GO:0006955	immune response	2.96E-8
GO:0050778	positive regulation of immune response	2.55E-6
GO:0045087	innate immune response	5.36E-6
GO:0002253	activation of immune response	6.33E-6
GO:0050776	regulation of immune response	2.01E-5
GO:0002684	positive regulation of immune system process	2.11E-5
GO:0034097	response to cytokine stimulus	3.12E-5
GO:0050853	B cell receptor signaling pathway	4.54E-5
GO:0006952	defense response	1.05E-4

Table 5.8.: GO enrichment analysis of 200 top ranked TB features. Shown are the 10 most significantly enriched GO terms.

drawn from a large AML data set comprising over 2000 samples from 17 different studies. Comparison with the prediction performance of the entire cohort reveals that initial pilot trials would require larger patient cohorts. To predict the minimum size of such a consecutive pivotal validation trial we established a method combining sample extrapolation and adaptive learning approaches. This strategy also estimates the overall best statistical performance to be expected from additional larger validation trials. Utilizing a combination of TSA and adaptive learning, we estimated trial size and test performance for pivotal trials using peripheral blood to develop three independent biomarker tests to detect patients with ATB, NSCLC and HIV. All approaches are predestined to significantly shorten the time from explorative pilot studies of signature-based biomarkers to their clinical routine use.

Conventional procedures often involve CV procedures to estimate the predictive performance of microarray classifiers. Early studies have already shown that for selecting a good classifier, 10-fold CV model selection gives more accurate results than the more expensive leave-one-out CV (*Simon et al.*, 2003). Although CV has been shown to give an unbiased estimate of prediction performance, the best model to estimate model's error rate is still controversial (*Molinaro et al.*, 2005; *Berrar et al.*, 2006; *Varma and Simon*, 2006). For small sample sizes it was shown that CV results in high variability of the error estimation and bootstrap methods provide improved performance, but often at the cost of increased bias (*Braga-Neto and Dougherty*, 2004). Bootstrap corrections with the .632+ bootstrap method are biased in small sample sizes with strong signal-to-noise ratios (*Molinaro et al.*, 2005). However, our intention in this study is not to compare several methods for estimating the 'true' prediction error of a prediction model but rather to mimic a typical clinical situation and to estimate the variance one obtains by subsequent patient recruitment into the different training and validation cohorts. TSA shows that for the small AML pilot trial, prediction performance is largely dependent on the TS composition and results based on the classical design of TS and VS should be interpreted carefully to avoid overoptimistic results. The overall development process may be inefficient if claims are based on initial, unreliable results (*Ransohoff*, 2007).

Ein-Dor et al. (*Ein-Dor et al.*, 2005) noticed by reanalyzing breast cancer prognostic data sets that many signatures with similar prediction accuracy can be developed from one data set. Although selecting a single optimized model is the most common approach for biomarker development, we and others (*Chen et al.*, 2011) analyzed whether the combination of different models could be preferable. Hence, we propose a feature selection method based on ranksum derived from different subsets of the original data set weighted by the contribution of each feature in the SVM. In this way, stable signatures with features present in different subsets of the original data sets are combined with classification performance. A similar approach was developed in (*Davis et al.*, 2006; *Abeel et al.*, 2010).

The number of features included in a final classifier is still controversial. Many authors argue that more features lead to overfitted results while other claim that modern classification techniques overcome these concerns (*Hua et al.*, 2005; *Dougherty and Brun*, 2006; *Sima and Dougherty*, 2006; *Simon*, 2008). We observe that less stringent features selection cut-offs lead to higher performance in TSA. On the other hand, classifiers with more features were not superior to random classifiers when applying classifiers from the large AML cohort to an independent validation set measured on a different platform (HG-U133 2.0), supporting the specificity of the classifiers with a smaller feature set.

ROC curves must be used with caution unless one has a very large sample (*Hanczar et al.*, 2010; *Berrar and Flach*, 2012). Although the AUC is the most frequently used method for performance assessment of a microarray classifier, one has to keep in mind that ROC analysis measures a model's ability to rank positive and negative cases relative to each other and that the AUC cannot directly measure clinical benefit or loss (*Hilden*, 2005). The partial AUC was proposed as an alternative measure for the full AUC (*Walter*, 2005). For completion, we included MCC, sensitivity and specificity as performance measurements with a threshold at the maximum Youden Index (*Youden*, 1950). Overall, lower prediction rates were obtained by the MCC may be due to a nonoptimal threshold. The choice of the decision threshold is not straightforward and depends on the particular clinical application, for example a screening diagnostic test might be optimized for high sensitivity (*Metz*, 1978; *Obuchowski*, 2005). Bootstrap validation by different subsets of the final validation set can present a range of possible performance independent of sample composition in a validation set as well.

Sample size estimation is an important aspect of microarray data analysis in order to achieve sufficient power with minimal sample size to reduce cost and time. Many methods are developed to determine the number of replicates required to have adequate power to identify differentially expressed genes (*Yang et al.*, 2003; *Wang and*

Chen, 2004; *van Iterson et al.*, 2009; *Lin et al.*, 2010), either based on theoretical considerations of the underlying distribution or based on pilot data. As differential expression was not the key objective of our project, we developed an approach applicable for classification analysis in a clinical setting. Our sample size estimation is intended to give clinicians guidance on patient numbers recruited for pivotal validation trials to obtain a specific prediction performance and therefore, we employed a data-driven approach based on simulated data. This simulated data is derived from a heuristic approach as we observed that a more sophisticated model based on the method of *Parrish et al.* (2009) resulted in overoptimistic prediction performance. A similar extrapolation approach was performed in the publication from Ein-Dor et al. modeling Gaussians distributions to generate simulated samples concluding that at least thousands of samples are needed to achieve stable feature selection (*Ein-Dor et al.*, 2006). Oversampling is also a common approach for learning from imbalanced data (*He*, 2009). We further observed that sample size estimation can slightly vary when re-analyzing existing data sets, indicating that information obtained from 100 different patient drawings together with 100 repetitions of TSA in each step may not be complete. I would envise enhancing this to 1000 different patient drawings. Still one should mention that for 100 permutations a 64-bit Windows system with 8GB of RAM needed approximatly 1 week for computation. Therefore, improving the existing scripts with for example using the multicore package in R (*Urbanke*, 2012) for running parallel computations should be considered to obtain computationally more efficient algorithms.

Finally, one should keep in mind that methods are developed based on a data set with good predictive performance, the AML data set. For a final proof of method, completely synthetic data should be generated, for example as described by *Molinaro et al.* (2005).

For the efficient translation of cancer genome information into the clinic, studies must go beyond statistical analyses of large genomic data sets. This process will require the inclusion of expertise and insights from molecular biology, genetics and systems biology, as well as clinical experiences. We performed functional analyses of resulting gene sets and observed disease-related enrichment of the obtained signatures and phenotype-related feature ranking. For example, the HIV signature is enriched for GO terms of viral infection which is in accordance with biological common sense and supports their functional relevance.

Taken together, the here proposed methods can further optimize the translation of the cancer genomic applications into effective diagnostic biomarkers. For AML diagnostics, a web-based prototype was already implemented with the intention of making the data and signature available for scientists not familiar with a command line based environment like the R language.

6. Critical consideration of a novel high-throughput technology

As shown in the preceding chapters, blood-based mRNA GEX profiling has become an important issue in medical applications and useful clinical predictions can be made from this high-throughput technology. In this section I would like to focus on the challenges when a novel high-throughput technology is applied to the development of blood-based diagnostic biomarkers. Reasonable concerns would arise from the reproducibility of expression profiles and the impact of RNA isolation methods. The following study interrogates the latter concern using a recently introduced bead-array based microRNA (miRNA) expression technology.

6.1. Motivation

In addition to blood-based GEX profiling, the characterization of the small RNA transcriptome including miRNAs has opened additional avenues for diagnostic approaches. MiRNAs are short (~22 nucleotides) non-coding RNAs and play a regulatory role by translational repression or degradation of specific target mRNAs at the post-transcriptional level (*Bartel*, 2004). MiRNA expression profiles have already been proposed as useful biomarkers in cancer and other diseases(*Calin et al.*, 2004, 2005; *Calin and Croce*, 2006; *Iorio et al.*, 2005; *Mattie et al.*, 2006; *Lawrie et al.*, 2007). One important result of these efforts is that expression profiling derived from miRNAs can be used for cancer classification with more accuracy than mRNA expression profiles *Lu et al.* (2005). Since miRNAs can be detected in blood and body fluids (*Skog et al.*, 2008), peripheral blood might be a perfect source to monitor tumor-associated miRNA expression signatures. Most recently, blood-based disease specific miRNA signatures were identified in lung cancer patients (*Keller et al.*, 2009a), patients suffering from multiple sclerosis (*Keller et al.*, 2009b) as well as in young stroke patients (*Tan et al.*, 2009). However, when applying miRNA expression profiling in routine clinical settings, the method of RNA preservation, the manner of RNA extraction as well as the reliability of the miRNA profiling procedure have to be carefully considered. Technical issues regarding high-throughput miRNA expression profiling are discussed by various authors (*Wang et al.*, 2008; *Nelson et al.*, 2008). More recently, a bead array based assay was introduced comprising of 735 human miRNAs allowing high-throughput expression profiling in a large number of samples (*Chen et al.*, 2008). In this project this newly introduced array platform is evaluated for miRNA expression profiling of peripheral blood. Our comprehensive miRNA screen is intended to serve as a reference for future studies assessing peripheral blood in the context of diagnosis and monitoring of certain diseases in clinical studies.

6.2. Results

First, the technical reproducibility of miRNA profiles was evaluated within and between different Sentrix Array Matrix (SAM) devices, which allow the assessment of 735 miRNA profiles of 96 samples in parallel. First, RNA samples derived from peripheral blood mononuclear cells (PBMC) of 11 different healthy donors (biological replicates) were analyzed in triplicates on one SAM (intra-SAM reproducibility) and in the next step the same 11 RNA samples used for the intra-SAM evaluation were analyzed on two additional SAMs (inter-SAM reproducibility). Before analyzing the data, quality assessment was performed and outliers were removed (see sec. C.2). Different normalization procedures were compared between replicate samples and as highest overall correlation between samples and the lowest variance was observed after quantile normalization (data not shown), this method was used in all following analyses. The intra- and inter-SAM reproducibility of the replicates was estimated by calculating the Pearson correlations (r^2) for all pair-wise combinations of individual miRNA profiles within a given sample. The overall mean correlation coefficient in intra-reproducibility was 0.9933 ± 0.0066. For inter-SAM analysis the overall mean correlation coefficient was 0.9880 ± 0.0069. Taken together, intra- and inter-SAM reproducibility of the miRNA microarray assay was very high (Fig. 6.2.1A).

To test the lowest amount of total RNA derived from peripheral blood mononuclear cells, that still yields reliable results in the miRNA assay, a titration experiment was performed using a broad range of input amounts of total RNA (2, 12.5, 25, 50, 100, 200, 500 ng) from three healthy individuals which were tested in triplicates. Analysis of variance was performed to assess the reproducibility of miRNA microarray data in the titration experiments by comparing all RNA input amounts below 500 ng against the 500 ng RNA reference and the correlation within each biological replicate were calculated. As demonstrated in Fig. 6.2.1B, correlations remained relative constant when using 100 to 500 ng total RNA and even 2 ng total RNA leads to reproducible ($r^2 > 0.9$) results. When comparing all input amounts of 2, 12.5, 25, 50, 100 and 200 ng to the reference of 500 ng there was a statistically significant difference (t-test p-value <0.05) for the samples with 50 ng or less input amount indicating that less than 100 ng total RNA is not sufficient to produce reliable results in peripheral blood derived samples.

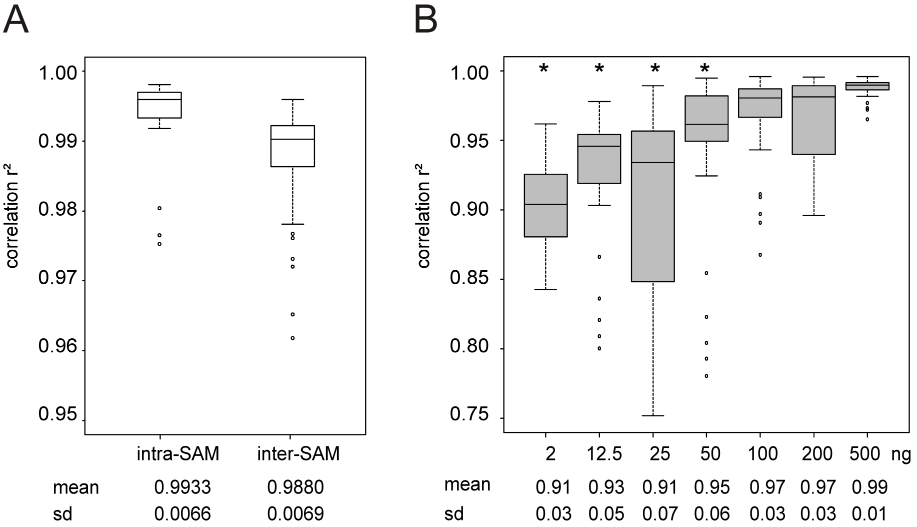


Figure 6.2.1.: Analysis of peripheral blood samples with miRNA microarrays A: The intra- and inter-array data reproducibility was estimated by calculating the Pearson correlation coefficients (r^2) for all pair-wise combinations of technical triplicates derived from 11 different blood donors analyzed on the same and on different arrays and displayed in boxplots. B: Performance of total RNA input was estimated by calculating r^2values between different input amount versus the 500 ng group. The boundary of the box closest to zero indicates the 25th percentile, the line within the box marks the median, and the boundary of the box farthest from zero indicates the 75th percentile. Whiskers above and below the box indicate the 90th and 10th percentiles. Outliers are plotted as dots. In B, * mark p-values < 0.01 (unpaired t-test) in comparison to the 500 ng group.

Next, we were interested if the presence of other RNA molecules such as mRNA including miRNA precursor RNA or ribosomal RNA had any influence on the blood expression profiling of miRNAs. To address this question, we compared PBMC derived miRNA profiles of total RNA to enriched low molecular weight (LMW) RNA profiles from six healthy individuals. Therefore, fractions of total RNA as well as of less abundant LMW enriched RNA below 200 bp were extracted. In contrast to the total RNA fraction, pri-miRNA sequences as well as mRNAs and ribosomal RNAs should be depleted in the LMW RNA fraction. By statistical testing for differentially expressed miRNAs between total and enriched LMW-RNA, 134 miRNAs out of 735 (18%) were identified to be significantly changed in the process of LMW RNA enrichment (FC > +/- 2, total difference in mean signal intensity between both groups > 100 and Benjamini-Hochberg adjusted p-value < 0.05). Unsupervised hierarchical clustering based on the most variable transcripts in this dataset demonstrate a clear separation of the expression profile between total and enriched small RNA (Fig. 6.2.2A). Taken together, there is a significant difference in the expression profiles between total and enriched LMW RNA indicating that the miRNA assay is sensitive to size fractionation.

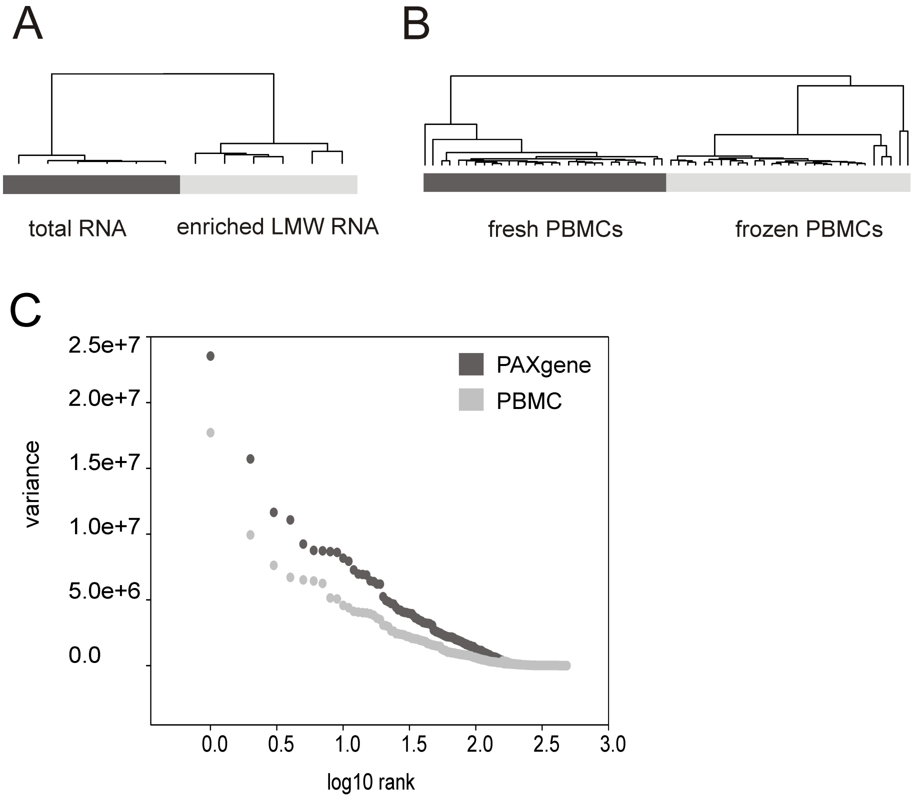


Figure 6.2.2.: Impact of RNA isolation and sampling handling on miRNA expression profile
Unsupervised hierarchical cluster analysis of (A) total and enriched LMW RNA and (B) freshly isolated or cryopreserved PBMC samples performed on quantile normalized data using the most variable miRNAs in the dataset (SD/mean 0.5 to 10). Only the cluster dendrogram is shown. C: Overall variance in PAXgene (dark gray) and PBMC (light gray) samples was ranked and plotted. Variance was calculated in quantile normalized data for each miRNA transcript that was called present (detection p-value <0.05) in at least one-third of the 12 samples. Statistical difference between corresponding whole blood and PBMC variance was calculated using paired t test and resulted in a p-value <0.05.

Since many tissues and cells from clinical samples are often stored for long time periods in liquid nitrogen, i.e. to allow retrospective gene or miRNA expression analysis, it is important to know if expression profiles remain stable or undergo in vitro changes. We therefore assessed whether the storage of viable cells in liquid nitrogen has any impact on the stability of miRNA transcripts. For this purpose, miRNA expression profiles derived from total RNA obtained from fresh or frozen PBMC of 29 healthy individuals were compared. Hierarchical cluster analysis clearly demonstrated an incisive effect on miRNA expression profiles, resulting in a clear separation of fresh PBMC from frozen specimens (Fig. 6.2.2B). None of the related sample pairs clustered together or were grouped together, indicating that fresh and frozen cells cannot be compared directly.

To determine the effect of different blood cell populations and their influence on miRNA expression profiling, we compared RNA derived from PBMC to RNA from whole blood samples. Therefore, blood samples from six donors were collected into either CPT tubes followed by PBMC separation and total RNA isolation using TRIZOL or were collected into PAXgene Blood RNA Tubes followed by direct RNA extraction using the newly introduced PreAnalytiX's PAXgene Blood miRNA Beta Version kit. There was no relevant difference in the amount and quality of the obtained RNA. When assessing miRNA microarray quality both RNA isolation techniques revealed comparable results. We observed increased variance in the whole blood samples (t-test p-value<0.05) in comparison to PBMC samples probably due to the increased heterogeneity of cell types in whole blood compared to the PBMC fraction (Fig. 6.2.2C). The clear differences between whole blood versus PBMC derived profiles lead to a total of 158 miRNAs (21.55%) differentially expressed miRNAs (p-value < 0.05, fold-change $> +/- 2$) and a mean r^2 between sources of 0.7889 $\pm$ 0.180.

To verify the accuracy of this bead-based technology and to validate miRNA expression, qPCR data was explored for 12 miRNAs for blood samples derived from the six aforementioned healthy donors. Three RNA isolation approaches were chosen: total RNA, enriched small RNA and PAXgene RNA. In all three comparisons absolute normalized expression values from the miRNA array were highly correlated to qPCR negative Ct values with Spearman's correlation r^2-values of 0.9597 for total RNA samples, 0.8346 for enriched small RNA samples and 0.8873 for PAXgene RNA samples (Fig. 6.2.3).

6.3. Summary and Discussion

To study the reproducibility of data generated on this new miRNA array platform we evaluated several technical replicates derived from peripheral blood samples of healthy subjects. A high reproducibility of miRNA expression data was estimated by calculating correlation coefficients of technical replicates on the same as well as on different SAM devices. All technical replicates analyzed revealed correlation coefficients > 0.98 indicating high reproducibility and reliability of this miRNA microarray assay. A high sensitivity of the miRNA assay method was shown for tissue samples and cell lines by Chen et al. (*Chen et al.*, 2008) and we extend these findings demonstrating that highly reproducible miRNA expression profiles are generated with 100 – 200 ng total RNA input from PBMC and stabilized whole blood.

In contrast to mRNA profiling technologies, miRNA profiling must take into account the difference between mature miRNAs and their precursors. The data of our study clearly indicate that assay results are different if using total RNA or

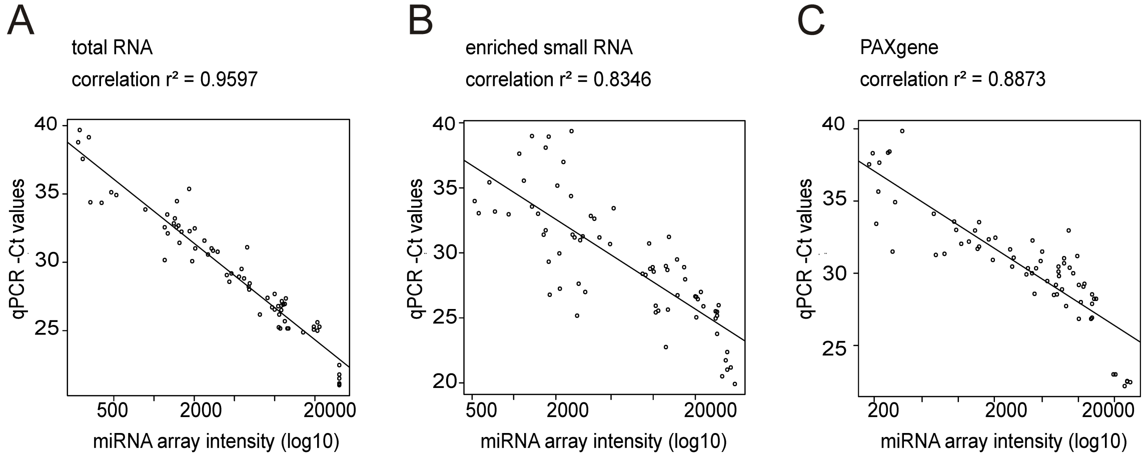


Figure 6.2.3.: Correlation of miRNA array and qPCR expression
Correlation between array and qPCR results are displayed in a scatterplot illustrating the association of normalized miRNA array expression (log10) versus the Ct value for six total RNA (A), enriched small RNA (B), and PAXgene samples (C) together with a linear regression fit line. Spearman's correlation coefficients were calculated as well for each RNA isolation technique.

enriched LMW-RNA derived from peripheral blood. However, reproducibility and correlation were very high if results were compared between assays using the same preprocessing procedures (r^2>0.99) showing that the enrichment procedure itself did not add variation to the measurement. Thus, enrichment of LMW-RNA can on one hand lower cross hybridization through inactive precursors of the miRNAs present in total RNA or through additional mRNA interference. On the other hand this procedure might also lead to a biased composition of mature miRNAs lacking for example less abundant active miRNA sequences. Both phenomena could account for the variations we observed when using RNA enrichment. However since total RNA as well as enriched LMW-RNA thereof give reliable and reproducible results both methods should be valid for diagnostic purposes when taken into account its specific limitations.

Cryopreservation of clinical specimens, including PBMC, is a commonly used technique in many clinical trials. In particular for multicenter studies, the performance of batch testing of harvested frozen samples in a central diagnostic unit is often preferred upon the independent analysis of fresh samples at different study sites. However, cryopreservation is associated with adverse effects on subsequent functional studies in comparison to those performed with freshly isolated cells. Here we estimated the influence of cryopreservation of isolated PBMC on miRNA expression profiles by comparing cryopreserved PBMC to their matched counterparts directly lysed after isolation. We clearly demonstrated that the miRNA profiles of cryopreserved PBMC samples were not comparable to freshly isolated PBMC similar to findings obtained in mRNA expression studies (*Debey et al.*, 2004; *Debey-Pascher et al.*, 2011). A careful and uniform selection of sample material therefore needs to be assured within a study and attention needs to be paid to those transcripts that are affected by sample handling procedures.

In a clinical setting peripheral blood is the most widely used tissue for disease monitoring. If miRNA expression profiling will become a routine tool for diagnostic and prognostic clinical studies it is crucial to understand the differences by using either whole blood samples or isolated PBMC. We compared miRNA expression profiles derived from PBMC using TRIZOL isolation and from whole blood using a PAXgene Blood miRNA Kit Beta Version extraction kit. Most importantly, we assessed the newly introduced PAXgene isolation method as a robust technique for total RNA including miRNA extraction of blood samples whose RNA quality is comparable to RNA obtained from PBMC using TRIZOL. Analysis of amount of present miRNAs, overall signal intensity and intra-group correlation gave similar results. Akin to mRNA expression profiling, whole blood derived miRNA profiles exhibit higher variance and results in miRNAs significantly differentially expressed in comparison to the PBMC derived profile.

Finally the expression of 12 miRNAs was determined by qPCR as another established expression technology. Good correlation coefficients between miRNA array expression levels and qPCR values were assessed for total RNA, enriched small RNA as previously reported (*Chen et al.*, 2008). PAXgene samples performed similarly well. We therefore conclude that the bead array platform revealed reliable results compared to the hitherto used qPCR method as golden standard for miRNA profiling.

Taken together, we present clear evidence that highly reproducible and reliable miRNA profiles of primary human peripheral mononuclear cells as well as of whole blood samples are obtained by the miRNA bead array technology. We clearly demonstrate that sample handling and the choice of either PBMC or whole blood is a rather critical issue when assessing miRNA profiles. Furthermore, we evaluated that the assay monitors different expression profiles when RNA size fractionation is performed. This comprehensive dataset of miRNA profiles derived from peripheral blood of healthy individuals might serve as a valuable resource for future steps towards the establishment of a comprehensive global miRNA profile of human peripheral blood for diagnostic as well as prognostic purposes.

7. Discussion and future perspectives

7.1. Discussion

More than 10 years after Golub et al. proposed for the first time that transcriptional-based classification can be used for the differential diagnosis of AML and ALL (*Golub et al.*, 1999), the number of tests applied to clinical practice is alarmingly small and to date no single test has been approved using gene expression profiling of peripheral blood (*Staratschek-Jox et al.*, 2009). Fundamental concerns regarding the microarray technology itself as well as the reliability of the technology in clinical applications such as disease diagnosis and prognosis were eased by results derived during the first and second phase of the MAQC project (*Shi et al.*, 2006; *Guo et al.*, 2006; *Kuo et al.*, 2006; *Canales et al.*, 2006; *Shippy et al.*, 2006; *Shi et al.*, 2010; *Luo et al.*, 2010; *Oberthuer et al.*, 2010; *Parry et al.*, 2010; *Huang et al.*, 2010; *Fan et al.*, 2010b). Hence, this thesis aims to accelerate the translation of basic findings to clinical utility. Several aspects in the process of diagnostic biomarker development are evaluated by analyzing genome-wide expression data derived from 4738 individuals in 37 studies addressing four clinical endpoints to assess over 1,800,000 different classifiers.

The initial hype of genome-wide technology has been replaced by realism, and it has become clear that many of the hopes based on early studies did not manifest due to underpowered studies, non-standardized sample procedures and insufficient bioinformatics analysis. However, blood-based transcriptional biomarkers have the potential to be used in routine clinical applications (*Baird*, 2006; *Staratschek-Jox et al.*, 2009). A good starting point will be diseases for which there is a lack of acceptable diagnostic tools and which would greatly benefit from an improvement. Hence, in this thesis, we focused on settings with (1) high enough prevalence and (2) a clinically relevant endpoint with a need for more relevant and reliable diagnostic tests. In our opinion, it is necessary to ensure that blood transcriptomics will not be purely seen as another add-on diagnostic method further increasing costs within our healthcare systems with its limited financial resources but rather that it will serve as a substitution technology for older, less reliable, and less informative technologies. In principle, genomic profiling could be used for primary diagnosis, differential diagnosis, subclassification of diseases, therapy outcome, prognosis, and prediction of diseases with one single technology. However, for example in leukemia diagnosis, transcriptional profiling has mostly been applied to optimize disease subclassification and prognostic as well as predictive purposes. We addressed this issue by re-evaluating the use of GEX for the primary diagnosis of AML. To answer this important question, we integrated data sets from 17 different studies to form a new data set comprising a total of 2013 samples. By comparing AML samples to a cohort of healthy controls as well as samples comprising other diseases not related to leukemia and other leukemia samples including ALL and chronic leukemia we addressed primary as well as differential diagnosis of acute leukemia. Furthermore, we also assessed AML subclassification based on cytogenetic aberrations in a first step towards a combined AML diagnostic tool.

While blood-based profiling is the obvious application for hematologic malignancies such as leukemia, an increasing number of studies assessed the use of blood transcriptomics for the diagnosis of other diseases as well (*Twine et al.*, 2003; *Burczynski et al.*, 2005; *Sharma et al.*, 2005; *Osman et al.*, 2006; *Critchley-Thorne et al.*, 2007; *Showe et al.*, 2009; *Staratschek-Jox et al.*, 2009; *Chaussabel et al.*, 2010). For many of these other disease states, the primary affected tissues are not readily available or can only be obtained through invasive intervention. Peripheral blood as a very promising alternative is an easily accessible and practical surrogate source (*Baird*, 2006; *Burczynski and Dorner*, 2006) based on the assumption that it reflects pathological changes because circulating blood is in contact with almost every organ and tissue in the body. In this thesis, we applied high-throughput microarray technology to answer the important clinical question of NSCLC, ATB, and HIV diagnosis. In all studies, the control cohort comprised healthy controls as well as closely related diseases. This study design ensures a higher potential of discovering the disease in a heterogeneous setting since variations in the selection of study populations will have a great impact on experimental results.

For all genomic studies to make an impact on medicine, a strict standardization of all procedures from sample handling to study design and bioinformatics analysis is crucial. Regarding sample handling, important protocols were established by the MAQC-I consortium providing quality control tools to avoid procedural failures (*Shi et al.*, 2006; *Guo et al.*, 2006; *Kuo et al.*, 2006; *Canales et al.*, 2006; *Shippy et al.*, 2006). Stabilization of RNA is one prerequisite for large multicenter studies, and RNA analysis of peripheral blood in particular relies on very careful sample procurement and processing (*Debey et al.*, 2004, 2006; *Debey-Pascher et al.*, 2011; *Elashoff et al.*, 2012). Furthermore, compatibility and therefore comparability between different microarray platforms (*Kuo et al.*, 2006; *Canales et al.*, 2006) as well as between older and new versions of microarrays (*Eggle et al.*, 2009; *Barbosa-Morais et al.*, 2010) is not straightforward and make the long-term development of biomarker diagnostics on this technology rather difficult. Hence, an important issue for the industry is to produce devices available over longer periods of time.

In addition to blood-based GEX profiling, the characterization of the small RNA transcriptome including miRNAs has opened additional avenues for diagnostic approaches (*Calin and Croce*, 2006). MiRNAs are small non-coding RNAs that play a role in transcriptional or post-transcriptional regulation of genes involved in numerous biological processes, including differentiation and proliferation. Profiling of miRNA expression levels has indicated that, similar to mRNA profiling, distinct miRNA signatures can be obtained. In further research, the platform requirements for blood-based biomarker development were exemplarily examined for miRNA expression profiling. Here, we could show that both intra- and inter-array reproducibility are key aspects when applying a novel high-throughput platform in a clinical setting. We have studied the performance of a miRNA expression profiling microarray platform when assessing samples derived from peripheral blood in healthy subjects. More importantly, we addressed reproducibility and sample handling issues as well as the influence of different blood-based RNA extraction methods.

Despite the MIAME standards defined in 2001 (*Brazma et al.*, 2001), data storage in databases such as GEO (*Edgar et al.*, 2002; *Barrett et al.*, 2011) and more importantly, comprehensive experimental annotation indicating clinical and epidemiological characteristics of data is still insufficient (*Kostka and Spang*, 2008; *Quackenbush*, 2009). We observed that when building the AML data set, downloading a data set from GEO is simple, but assigning respective class labels often has to be done manually by browsing through the related article and its given supplement. Often, the annotation of samples is incomplete, and detailed clinical documentation that allows a comparison to be made for the different studies was not given, rendering large-scale meta-analysis and integration of GEX data a challenging task. To make data available beyond their initial publication, effective data-reporting standards should be enforced.

Another major requirement in order to determine early study limitations is the standardization of bioinformatics data analysis. As suggested by the MAQC-II study, we first established our preferred standard classification process comprising feature selection, algorithm and parameter optimization in internal 10-fold CV repeated 10 times and performance assessment through ROC analysis as discussed in chapter 4. All bioinformatics approaches are documented in supplemental R scripts making analysis procedures available to and reproducible by others (Appendix D). Using this standard approach, a RNA-stabilized whole blood diagnostic classifier for NSCLC consisting of 484 transcripts could be established in a TS that can be used as a biomarker to discriminate between NSCLC patients and control samples. Subsequently, this optimized classifier was successfully applied to two independent and blinded VSs resulting in AUC values of 0.824 (VS1) and 0.977 (VS2). This successful application of the classifier in both VSs underlines the validity and robustness of the NSCLC classifier. Extensive permutation analysis using random feature lists supports the specificity of the established transcriptional classifier. We further showed that the use of this approach to make clinical predictions performs similarly well to best-performing models obtained in the MAQC-II Multiple Myeloma data set analyzing 4 different endpoints.

The complexity and volume of transcriptional profiling requires heightened attention to robust design, methodological details and avoidance of bias to making genomic-based biomarker useful for the clinic (*Ioannidis*, 2007). Insufficient sample size resulting in lower power of GEX studies is a major drawback in the process of translating blood transcriptomics into clinical routine practice. To analyze the validation of claims derived from initial pilot studies, we developed TSA as a method to investigate the overall classifier performance that can be expected independent of the actual clinical situation with subsequent patient recruitment into training and validation cohorts. We used a small pilot trial of 150 samples randomly drawn from the large AML data set to mimic a typical clinical trial. Using this PTS, the variance obtained by random assignments into the cohorts was estimated for different feature selection and classification algorithms as well as for different feature selection cut-offs. Linear SVM classification in combination with a t-test outperformed other methods and was subsequently used for all upstream analyses. We could show that prediction performance is largely dependend on sample cohort inclusion. Furthermore, we demonstrated that many models showing good performance in classifying the two VSs when using the conservative workflow of one single division into data sets did not reach sufficient statistical performance when applied to the remaining samples of the large AML cohort. Hence, we propose a larger sample-split approach to investigate the overall variance of prediction performance and to predict the outcome of a large biomarker validation study from the PTS.

Next, we assessed the predictive performance obtained when using the large AML cohort of 2013 samples. We proved that in this data set, a robust GEX-based classifier as a test for the primary diagnosis of AML can be developed with a high sensitivity and specificity. In comparison to the small PTS, all random drawings of cohorts reached AUC values >0.977 and 63% came to an AUC>0.99 whereas in the small PTS, only 2.4% resulted in such a high AUC. Feature lists were rather stable in terms of feature sizes derived from different cut-offs and the individual transcripts included in signatures. These results also indicate that even initial PTS would require larger patient cohorts for a robust classifier development, a requirement that can rarely be met. The overall development process may be inefficient if claims are based on initial, unreliable results (*Ransohoff*, 2007).

Methods to select a good classifier from the PTS were compared by different ways to agglomerate feature lists obtained during TSA. A method averaging feature weights for SVM classification performed best in this setting in comparison to both methods simply using all DE signatures without summarized SVM weights and methods ranking features according to FC and p-value. Hence, we propose a feature selection method based on ranksum derived from different subsets of the original data set weighted by the contribution of each feature. In this way, stable signatures with features present in different subsets of the original data sets are combined with classification performance. Ensemble feature selection with data perturbation of random subsets has also been proposed by others (*Abeel et al.*, 2010; *Davis et al.*, 2006) and are discussed in *He and Yu* (2010). The stability of feature selection can also be analyzed as described by Boulesteix and Slawiski (*Boulesteix and Slawski*, 2009), with the Davis index (*Davis et al.*, 2006) or with the Kuncheva index (*Kuncheva*, 2007).

We then applied our approach in order to build a final biomarker for AML primary diagnosis using all samples in the large AML cohort (n=2013). This classifier was applied to an independent and blinded AML data set of additional 2088 samples not used in any previous analysis, analyzed on a newer array version than the original AML data set (HG-U133 2.0). High predictive performance was derived in this VS (AUC>0.97). Classifiers with few features (50 features or less) were also significantly superior to 1000 random classifiers supporting the specificity of the established signatures. More sophisticated methods for batch effect removal when comparing two data sets performed on a different platform such as the ComBat algorithm (*Johnson et al.*, 2007) might even lead to a higher validation performance. I also developed a web-based AML diagnostic system prototype for scientists not familiar with a command line environment like the R language.

Completely specified and analytically validated biomarker candidates may not be feasible before its use in pivotal clinical trials and adaptive trial designs have been proposed for phase III trials (*Freidlin and Simon*, 2005; *Jiang et al.*, 2007; *Simon*, 2008). Consequently, we developed an adaptive learning approach in contrast to a static use of initial best-performing classifiers for (a) a continuous classifier optimization and (b) an estimation of the sample size required for a larger pivotal validation cohort. In essence, continuous classifier optimization can be achieved by adaptive learning techniques opening completely novel opportunities when integrating large parallel biological data for medical diagnostics and curve sketching of adaptive learning results allowed the calculation of the number of patients required to reach a certain level of statistical performance in such a pivotal clinical trial.

Sample generation is often not possible or time-consuming due to high costs and low prevalence of certain diseases. This practical limitation may be solved by the inclusion of artificial samples that can be simulated according to the distribution of given samples. Hence, we investigated the use of oversampling methods to improve classifier development from a small PTS. Sample simulation based on a given data set could in principle be used for three approaches: to improve classification performance when using more samples, to obtain more stabilized feature ranking, or to make predictions on a test set. By comparing sample size estimations from the large and extrapolated AML data set, we could show that an estimate of the sample size from the extrapolated data can be obtained, although these estimates are higher than those from the original cohort. We could further demonstrate that although the large and extrapolated AML data set showed a high concordance in terms of feature stability, feature ranking and prediction performance within each data set, the signatures stabilized by extrapolation did not result in enhanced prediction performance.

Utilizing the established approaches TSA and adaptive learning, we estimated the trial size and test performance for pivotal trials using peripheral blood to develop three independent tests to detect patients with NSCLC, HIV-positive patients, and patients with ATB. All signatures obtained by weighted SVM feature ranking showed a disease-related pathophysiology and phenotype-related feature ranking. The AML signature comprised a significant Gene Ontology enrichment for immune system processes as well as related pathways such as T-cell, lymphocyte and leukocyte activation. Furthermore, leukemia was the most relevant significantly enriched Disease Ontology term. The NSCLC signature revealed an association with immune and immune system processes while the TB signature was enriched for immune response, cytokine stimulus, and B cell receptor signaling pathway. Interestingly, enriched GO terms of the blood-based HIV signature included viral transcription and viral infectious cycle supporting the functional relevance of the obtained signatures. Therefore, signatures also fulfilled the desirable biomarker characteristic of accuracy, stability, and mechanism (*Shi et al.*, 2010; *Fan et al.*, 2010b).

Thus in summary, using several simple clinical questions as models novel strategies that are aimed at closing the gap from initial promising pilot studies to clinical application of novel biomarkers were developed. I envision the following strategy for GEX classifier development:

1. TSA with 1000 permutations or more to estimate variance of prediction performance from a pilot study
2. Agglomeration of features by weighted SVM ranking
3. Random classifiers as controls to analyze specificity of established signature
4. Adaptive learning of extrapolated data to estimate sample size required for a larger pivotal validation cohort

7.2. Perspectives for future research

The here presented methods and results open several perspectives for future research. First, more samples should be recruited to improve the proposed signatures for NSCLC, HIV-positive patients, and patients with ATB according to sample sizes estimated by adaptive learning. Control cohorts should include healthy controls as well as closely related diseases to improve the study design. Heighted attention should be paid to covariates such as age and gender. For example, a larger control cohort for NSCLC should also comprise SCLC to ensure differential diagnosis of lung cancer and smoking levels should be comparable in cases and respective controls.

Next, the decision threshold currently taken at the maximum YI should also be optimized. This decision rule for classifying the test results as either positive or negative should be set according to the specific clinical application, for example a screening diagnostic test can be optimized for high sensitivity (*Metz*, 1978; *Obuchowski*, 2005). The threshold selection should be analyzed in the given data set and can then be used to predict new samples. A proper screening program improves patients' outcome by early detection and minimizes possible burdens from false-positive results.

Another perspective for future research would be the evaluation the number of features included in a final biomarker set, as the feature size is still controversial (*Hua* *et al.*, 2005; *Dougherty and Brun*, 2006; *Sima and Dougherty*, 2006; *Simon*, 2008; *Chang et al.*, 2011). Random signatures should be generated to verify the specificity of the established classifier. Recursive feature elimination as proposed by (*Guyon et al.*, 2002; *Duan et al.*, 2005; *Zhang et al.*, 2006; *Mundra and Rajapakse*, 2010; *Li et al.*, 2012) can be applied to identify the best-performing set of features.

One limitation of the current setting is that all classification models were applied to two-class diagnostic problems. Hence, the next steps would be expanding to multiclass classification tasks, for example separating HIV+, LTNP, and control samples or separating ATB, LTB, and controls. Multiclass classification algorithms include fuzzy SVMs and decision tree approaches (*Lee and Lee*, 2003; *Nguyen and Rocke*, 2002; *Tsujinishi and Abe*, 2003; *Mao et al.*, 2005) and a comprehensive evaluation of these multicategory classification methods have been performed by *Statnikov et al.* (2005) and *Li et al.* (2004). As a first step, we already compared different multiclassification algorithms for blood-based prediction analysis of different diseases (*Henseler*, 2009). The assessment of the prediction performance for such a case needs to be defined. Multi-class ROC graphs and AUCs are more complex since the entire space needs to be managed. With n classes the confusion matrix becomes an $n \times n$ matrix with n correct classifications on the diagonal and $n^2 - n$ possible errors on the off-diagonal (*Hand and Till*, 2001).

Additionally, one should keep in mind that methods are developed based on a data set with good predictive performance, the AML data set. For a final proof of method, completely synthetic data should be generated, for example as described in *Molinaro et al.* (2005) or *Van den Bulcke et al.* (2006) and TSA as well as adaptive learning should be repeated using this data as well. This data also opens the possibility to confirm claims of prediction performance by TSA.

Over the last years, an increasing number of other high-throughput technologies apart from GEX were established, and the generation of so called 'omics' data is commonplace in various biomedical research fields. Interesting novel options for clinical applications offer the profiling of miRNA levels, chromosomal copy number changes, epigenetic modifications, DNA sequencing and protein levels. Meta-analysis by combining multiple studies for a conclusive finding has become popular (*Tseng et al.*, 2012) and the inclusion of other data such as pathway information (*Ramilo et al.*, 2007; *Chang and Ramoni*, 2009; *Li et al.*, 2009; *Ma and Kosorok*, 2010) or protein-protein interaction networks (*Chuang et al.*, 2007; *Taylor et al.*, 2009; *Cun and Frohlich*, 2012) might enhance its predictive power.

Recently, declining costs for novel high-throughput sequencing platforms ("next-generation" sequencing) made this technology available for the application to larger clinical studies (*Shendure and Ji*, 2008; *Ding et al.*, 2010; *Cronin and Ross*, 2011; *Su et al.*, 2011). High-throughput sequencing is in particular attractive for clinical utility as the genome coverage is less biased and the dynamic range is larger (*Schuster*, 2008). Still, standards already available for sample processing, data storage, and especially data analysis need to be met for clinical translation. Similar to MAQC-

I and MAQC-II, the third phase of the MAQC project (MAQC-III), also called Sequencing Quality Control (SEQC), aims at assessing the technical performance of next-generation sequencing platforms and will evaluate various bioinformatics strategies as well.

Complementary technologies may be successful for endpoints not predictable by gene expression profiling alone and the most effective technology may also vary by study endpoint (*Tillinghast*, 2010). All approaches presented in this thesis can be applied to other technologies and help in bringing these technologies toward the clinic as well. There is a need for bioinformatics to handle the massive amount of data produced and to develop novel methods allowing the integration of various kinds of data, thus providing challenging opportunities for future research. Careful planning and robust study design is required to incorporate molecular profiling into clinical practice and we propose our strategy to be extended to other high-throughput data.

A. Supplementary tables

Cytogenetic subgroup	Number of samples	Risk group assignment
normal karyotype	336	favourable or intermediate
t(15;17)	43	favourable
t(8;21)	41	favourable
inv(16)	41	favourable
t(11q23)	44	unfavourable or intermediate
complex other	25	
single aberrations	54	
other	63	
NA	114	

Table A.1.: AML subclassification based on cytogenetic aberrations. Risk group assignment is taken from *Chen et al.* (2010).

	subgroup	total number	female	male	n.d.	mean age
HIV	HIV	79	13	50	16	40
	LTNP	27	10	16	1	41
controls	healthy controls	88	44	38	6	43
	sepsis	34	16	18	0	56
	sclerodemia	29	23	6	0	62

Table A.2.: Clinical and epidemiological characteristics of cases with HIV and respective controls; mean age is given in years. n.d. not defined.

group	total number	female	male	mean age
ATB	54	19	35	37
LTB	69	40	29	29
controls	24	15	9	30

Table A.3.: Clinical and epidemiological characteristics of cases with TB and respective controls; mean age is given in years.

	AUC	MCC	Sensitivity	Specificity
10 highest ranked features	0.9671	0.8587	0.9561	0.8995
20 highest ranked features	0.9785	0.8733	0.9469	0.9256
30 highest ranked features	0.9762	0.8536	0.8902	0.96382
50 highest ranked features	0.9964	0.9540	0.9716	0.9829
75 highest ranked features	0.9853	0.9322	0.9433	0.9899
100 highest ranked features	0.9818	0.9193	0.9323	0.9879
150 highest ranked features	0.9932	0.9636	0.9762	0.9879
200 highest ranked features	0.9972	0.9684	0.9927	0.9749

Table A.4.: Large AML classifier performance on HG-U133 2.0 AML data set
Shown are AUC, MCC, sensitivity and specificity for the 10, 20, 30, 50, 75, 100, 150 and 200 highest ranked features from feature ranking in the large cohort to build a classifier using the large HG-U133A data applied to the HG-U133 2.0 data set. MCC, sensitivity and specificity were assessed at the maximum Youden Index.

	AUC	MCC	Sensitivity	Specificity
10 highest ranked features	0.9050	0.6893	0.9378	0.7327
20 highest ranked features	0.9181	0.6927	0.9341	0.7417
30 highest ranked features	0.9270	0.7036	0.9360	0.7518
50 highest ranked features	0.9542	0.7494	0.8317	0.9176
75 highest ranked features	0.9486	0.7244	0.8417	0.8834
100 highest ranked features	0.9513	0.7537	0.8070	0.9437
150 highest ranked features	0.9727	0.8363	0.8756	0.9608
200 highest ranked features	0.9750	0.8425	0.8820	0.9608

Table A.5.: Small AML classifier performance on HG-U133 2.0 AML data set
Shown are AUC, MCC, sensitivity and specificity for the 10, 20, 30, 50, 75, 100, 150 and 200 highest ranked features from feature ranking in the small cohort to build a classifier using the small HG-U133A data applied to the HG-U133 2.0 data set. MCC, sensitivity and specificity were assessed at the maximum Youden Index.

Table A.6.: Annotation of AML signature

Rank	Probe ID	Gene Symbol	Entrez ID	RefSeq ID
1	207094_at	CXCR1	3577	NM_000634
2	203435_s_at	MME	4311	NM_000902
3	217023_x_at	TPSAB1	64499	NM_003294
4	204561_x_at	APOC2	344	NM_000483
5	204007_at	FCGR3B	2215	NM_000570
6	209995_s_at	TCL1A	8115	NM_001098725
7	210549_s_at	CCL23	6368	NM_005064
8	201427_s_at	SEPP1	6414	NM_001085486
9	206622_at	TRH	7200	NM_007117
10	216474_x_at	TPSAB1	64499	NM_003294
11	205051_s_at	KIT	3815	NM_000222
12	210084_x_at	TPSAB1	7177	NM_003294
13	209905_at	HOXA9	3205	NM_152739
14	210321_at	GZMH	2999	NM_033423
15	39318_at	TCL1A	8115	NM_001098725
16	203948_s_at	MPO	4353	NM_000250
17	203434_s_at	MME	4311	NM_000902
18	215382_x_at	TPSAB1	7177	NM_003294
19	204006_s_at	FCGR3A	2214	NM_000569
20	214651_s_at	HOXA9	3205	NM_152739
21	210119_at	KCNJ15	3772	NM_002243
22	205683_x_at	TPSAB1	7177	NM_003294
23	216782_at	—	—	—
24	211163_s_at	TNFRSF10C	8794	NM_003841
25	204885_s_at	MSLN	10232	NM_001177355
26	203828_s_at	IL32	9235	NM_001012631
27	221345_at	FFAR2	2867	NM_005306
28	207907_at	TNFSF14	8740	NM_003807
29	220068_at	VPREB3	29802	NM_013378
30	203691_at	PI3	5266	NM_002638
31	207826_s_at	ID3	3399	NM_002167
32	207741_x_at	TPSAB1	7177	NM_003294
33	221558_s_at	LEF1	51176	NM_001130713
34	205366_s_at	HOXB6	3216	NM_018952
35	207134_x_at	TPSB2	64499	NM_024164
36	207008_at	CXCR2	3579	NM_001168298
37	205568_at	AQP9	366	NM_020980
38	203949_at	MPO	4353	NM_000250
39	213110_s_at	COL4A5	1287	NM_000495

40	205131_x_at	CLEC11A	6320	NM_002975
41	210998_s_at	HGF	3082	NM_000601
42	220010_at	KCNE1L	23630	NM_012282
43	206135_at	ST18	9705	NM_014682
44	205798_at	IL7R	3575	NM_002185
45	214575_s_at	AZU1	566	NM_001700
46	209670_at	TRAC	28755	—
47	204891_s_at	LCK	3932	NM_001042771
48	204115_at	GNG11	2791	NM_004126
49	204698_at	ISG20	3669	NM_002201
50	210997_at	HGF	3082	NM_000601

Table A.7.: Annotation of NSCLC signature

Rank	Probe ID	RefSeq ID	Entrez ID	Gene Symbol	Illumina ID
1	10279	NM_005621	6283	S100A12	ILMN_1748915
2	2810040	NM_145699	200315	APOBEC3A	ILMN_1680192
3	1580259	XM_497072	389787	LOC389787	ILMN_1665823
4	2370524	NM_007115	7130	TNFAIP6	ILMN_1785732
5	4390242	NM_004084	1667	DEFA1	ILMN_1679357
6	6980537		NA		ILMN_1905548
7	3400551	NM_006138	932	MS4A3	ILMN_1726552
8	4060066	NM_000419	3674	ITGA2B	ILMN_1747248
9	4560133	NM_005139	306	ANXA3	ILMN_1694548
10	360066	NM_013378	29802	VPREB3	ILMN_1700147
11	6960440	NM_001925	1669	DEFA4	ILMN_1753347
12	4050286	XM_928682	645671	LOC645671	ILMN_1710007
13	6860754	NM_000045	383	ARG1	ILMN_1812281
14	1190349	NM_002759	5610	EIF2AK2	ILMN_1706502
15	70338	NM_004510	3431	SP110	ILMN_1672661
16	6350364	NM_002704	5473	PPBP	ILMN_1767281
17	5900072	XM_937928	347376	LOC347376	ILMN_1704385
18	1260228	NM_021105	5359	PLSCR1	ILMN_1752889
19	6180161	XM_371741	389293	LOC389293	ILMN_1675421
20	990097	NM_001816	1088	CEACAM8	ILMN_1806056
21	5080398	NM_003263	7096	TLR1	ILMN_1731048
22	520228	NM_182697	7328	UBE2H	ILMN_1757644
23	6400736	NM_004345	820	CAMP	ILMN_1688580
24	7570079	NM_002185	3575	IL7R	ILMN_1675949
25	2340110	NM_032321	84281	MGC13057	ILMN_1809636
26	580307	NM_138444	115207	KCTD12	ILMN_1742332
27	1450309	NM_002934	6036	RNASE2	ILMN_1730628
28	520646	NM_000713	645	BLVRB	ILMN_1797793
29	7400097	NM_001062	6947	TCN1	ILMN_1768469
30	6760255	NM_000104	1545	CYP1B1	ILMN_1693338
31	620324	XM_936731	647673	LOC647673	ILMN_1757702
32	7160608	NM_005067.5	6478	SIAH2	ILMN_1801313
33	430328	NM_016633	51327	ERAF	ILMN_1696512
34	2680273	NM_004926	677	ZFP36L1	ILMN_1675448
35	1090427	XM_928349	653600	LOC653600	ILMN_1693262
36	7650678	NM_017709	54855	FAM46C	ILMN_1713266
37	4210414	NR_002204	2503	FTHL11	ILMN_1706013
38	2760463	XM_931683	389293	LOC389293	ILMN_1672024
39	6960554	NM_005564	3934	LCN2	ILMN_1692223

40	1240044	NM_002483	4680	CEACAM6	ILMN_1712522
41	5570484	NM_005143	3240	HP	ILMN_1812433
42	3170241	NM_000140	2235	FECH	ILMN_1774091
43	610148	NM_001725	671	BPI	ILMN_1766736
44	160348	NM_002935	6037	RNASE3	ILMN_1802867
45	6200221	NM_021083	7504	XK	ILMN_1759117
46	4120270	NM_018566	55432	YOD1	ILMN_1678919
47	7560072	NM_206962	3275	PRMT2	ILMN_1748922
48	2650440	NR_002200	2497	FTHL2	ILMN_1746525
49	3830138	NM_020841	114882	OSBPL8	ILMN_1782459
50	4180564	XM_371243	388621	LOC388621	ILMN_1677262

Table A.8.: Annotation of HIV signature

Rank	Probe ID	RefSeq ID	Entrez ID	Gene Symbol	Illumina ID
1	6510553	NM_001768	925	CD8A	ILMN_1768482
2	4050703		NA		ILMN_1881490
3	1850047	NM_206833	404217	CTXN1	ILMN_1759766
4	5550364	XR_000904	144581	RPL14L	ILMN_1769937
5	1660021	XM_942993	647741	LOC647741	ILMN_1795159
6	940524	NM_005731	10109	ARPC2	ILMN_1810200
7	4390242	NM_004084	1667	DEFA1	ILMN_1679357
8	630022	NM_000683	152	ADRA2C	ILMN_1733963
9	4850128	NM_033423	2999	GZMH	ILMN_1731233
10	6560114	NM_001003	6176	RPLP1	ILMN_1689725
11	5270100	XM_943751	649516	LOC649516	ILMN_1661566
12	580131		NA		ILMN_1862684
13	5700687	NM_016024	51634	RBMX2	ILMN_1678203
14	2340379	NM_006917	6258	RXRG	ILMN_1750624
15	7610608	NM_020439	57172	CAMK1G	ILMN_1804339
16	3360603	NM_015229	23277	KIAA0664	ILMN_1770719
17	3120020	NM_002286	3902	LAG3	ILMN_1813338
18	4150546	XM_931897	653210	LOC653210	ILMN_1651606
19	540470	NM_018090	55707	NECAP2	ILMN_1749011
20	3780114	NM_001004708	219983	OR4D6	ILMN_1813776
21	6520497	NM_001614	71	ACTG1	ILMN_1704961
22	2320504		NA		ILMN_1915345
23	5050553	NM_014219	3803	KIR2DL2	ILMN_1678882
24	540731	NM_001768	925	CD8A	ILMN_1760374
25	4070129	NM_005091	8993	PGLYRP1	ILMN_1704870
26	6200551		NA		ILMN_1892403
27	5700541	NM_001013704	440313	LOC440313	ILMN_1749984
28	2100687	XM_928457	645416	LOC645416	ILMN_1700998
29	7550273	XM_931729	643665	LOC643665	ILMN_1664118
30	3840050	XM_495863	387751	GVIN1	ILMN_1668526
31	3310091	NM_005217	1668	DEFA3	ILMN_1725661
32	2630647	NM_207378	388007	SERPINA13	ILMN_1743046
33	3180075	NM_001008	6192	RPS4Y1	ILMN_1783142
34	1570491	XM_930555	653757	LOC653757	ILMN_1693207
35	2060706	NM_173574	257101	ZNF683	ILMN_1678238
36	2320400	XM_932788	645284	LOC645284	ILMN_1705982
37	3840524	NM_173566	253143	C22orf30	ILMN_1741295
38	4290692	NM_171825	815	CAMK2A	ILMN_1666445
39	1940414	NM_018264	55253	TYW1	ILMN_1736135

40	630274		NA		ILMN_1823884
41	6130128	XM_943174	648003	LOC648003	ILMN_1725702
42	4200671	NM_017409	3226	HOXC10	ILMN_1725899
43	4010707	NM_001003945	210	ALAD	ILMN_1679898
44	7550504	NM_033101	85329	LGALS12	ILMN_1776283
45	4200685	NM_003970	9172	MYOM2	ILMN_1716733
46	6980164	NM_000478	249	ALPL	ILMN_1701603
47	2710674	NM_012418	25794	FSCN2	ILMN_1795472
48	4250037	NM_178865	347735	SERINC2	ILMN_1694509
49	1050475	NM_014513	3810	KIR2DS5	ILMN_1691803
50	4260386	NM_153231	162972	ZNF550	ILMN_1760102

Table A.9.: Annotation of TB signature

Rank	Illumina ID	RefSeq ID	Entrez ID	Gene Symbol	Probe ID
1	ILMN_2114568	NM_052942	115362	GBP5	1510364
2	ILMN_2261600	NM_001017986	2210	FCGR1B	2710709
3	ILMN_2391051	NM_001004340	2210	FCGR1B	6620209
4	ILMN_2302757	NM_003890	8857	FCGBP	130609
5	ILMN_1809467	NM_006634	10791	VAMP5	2630195
6	ILMN_2394210	NM_138718	116369	SLC26A8	5360626
7	ILMN_1776939	NM_152866	931	MS4A1	3190521
8	ILMN_1851599		NA		3930128
9	ILMN_1662731	NM_031365	1285	COL4A3	6380066
10	ILMN_2388547	NM_033255	94240	EPSTI1	5700725
11	ILMN_1805750	NM_021034	10410	IFITM3	6650242
12	ILMN_1740572	NM_000355	6948	TCN2	5670100
13	ILMN_1664330	NM_001712	634	CEACAM1	5340767
14	ILMN_1701114	NM_002053	2633	GBP1	6840035
15	ILMN_1670305	NM_001032295	710	SERPING1	2030309
16	ILMN_1727271	NM_173701	7453	WARS	3710068
17	ILMN_1690241	NM_138456	116071	BATF2	4730059
18	ILMN_1755843	NM_052961	116369	SLC26A8	270240
19	ILMN_2176063	NM_000566	2209	FCGR1A	520086
20	ILMN_1675756	NM_170736	3772	KCNJ15	1050215
21	ILMN_2408987	NM_001003802	6604	SMARCD3	460463
22	ILMN_1701621	NM_005138	9997	SCO2	7510537
23	ILMN_2374865	NM_001040619	467	ATF3	4780128
24	ILMN_1690105	NM_007315	6772	STAT1	1820750
25	ILMN_1671703	NM_001613	59	ACTA2	6480059
26	ILMN_1667114	NR_003662	388524	LOC388524	460050
27	ILMN_1654389	XM_001128342	728744	LOC728744	5570039
28	ILMN_1691071	NM_032738	84824	FCRLA	3780193
29	ILMN_1772466	NM_005490	10045	SH2D3A	5490452
30	ILMN_1810289	NM_133337	26509	FER1L3	3170273
31	ILMN_1759075	NM_012452	23495	TNFRSF13B	620484
32	ILMN_2049184	NM_004944	1776	DNASE1L3	290739
33	ILMN_1688631	NM_145343	8542	APOL1	6370470
34	ILMN_1887868		NA		5270403
35	ILMN_2148785	NM_002053	2633	GBP1	2190148
36	ILMN_1771385	NM_052941	115361	GBP4	1980524
37	ILMN_1701455	NM_018438	26270	FBXO6	3800398
38	ILMN_2390946	NM_145350	8578	SCARF1	5390204
39	ILMN_1673503	NM_001005747	785	CACNB4	2260446

40	ILMN_1707979	NM_001007232	440068	INCA	3930368
41	ILMN_1780831	NM_003044	6539	SLC6A12	6220332
42	ILMN_2279844	NM_005356	3932	LCK	130274
43	ILMN_1733998	NM_005771	10170	DHRS9	630315
44	ILMN_1756953	NM_198460	163351	GBP6	3780047
45	ILMN_2337655	NM_004184	7453	WARS	4860224
46	ILMN_1747744	NM_005779	10184	LHFPL2	360132
47	ILMN_1683792	NM_015907	51056	LAP3	3290292
48	ILMN_1700671	NM_016135	51513	ETV7	6370768
49	ILMN_1701914	NM_014143	29126	CD274	4900239
50	ILMN_1669497	NM_017784	114884	OSBPL10	6760593

Disease	Cohort	Condition	n	Summe
AML	small	Feature selection	6	60,000
		Algorithm	1	80,000
		TS size		80,000
	large	Feature selection	6	6,000
		Algorithm	1	8,000
		ALA	32	320,000
	simulated	Feature selection	6	6,000
		ALA	32	320,000
TB	original	Feature selection	5	5,000
	simulated	Feature selection	5	5,000
		ALA	29	290,000
NSCLC	original	Feature selection	5	5,000
	simulated	Feature selection	5	5,000
		ALA	30	300,000
HIV	original	Feature selection	5	5,000
	simulated	Feature selection	5	5,000
		ALA	30	300,000
$\sum$				1,800,000

Table A.10.: Overview of the total number of classifiers established for trial simulation and adaptive learning assessment.

B. Supplementary figures

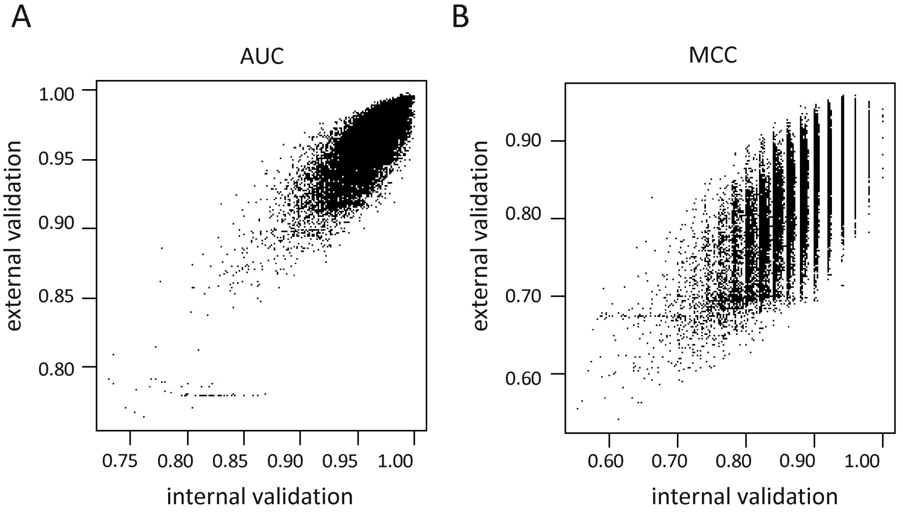


Figure B.0.2.: Model performance on internal validation compared with external validation for AML data
Shown is the (A) AUC and (B) MCC performance of 60,000 models derived from the small AML pilot trial in TSA (internal validation) plotted against the performance of each model build in TSA applied to the remaining blinded samples from the large cohort (external validation). Corresponding r^2-Pearson Correlation Coefficiants are 0.7976 (AUC) and 0.7294 (MCC).

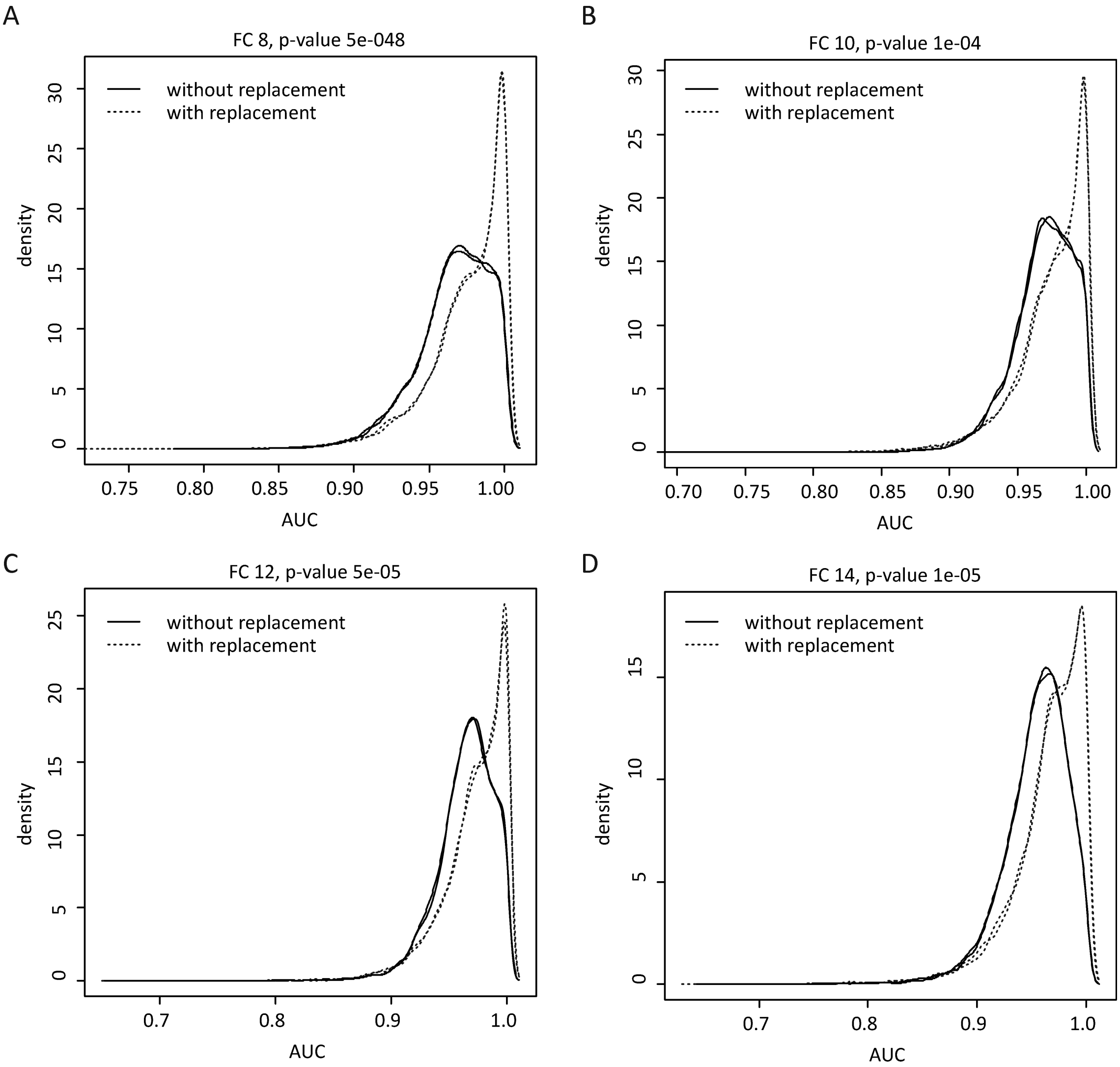


Figure B.0.3.: Sample-split versus bootstrap predictions for small AML cohort Shown are AUC distributions for 10,000 iterations of TSA for different FC/p-value filter for sample split (sample drawing without replacement, black lines) and classical bootstrap (sample drawing with replacement, dashed lines), displayed as density plots.

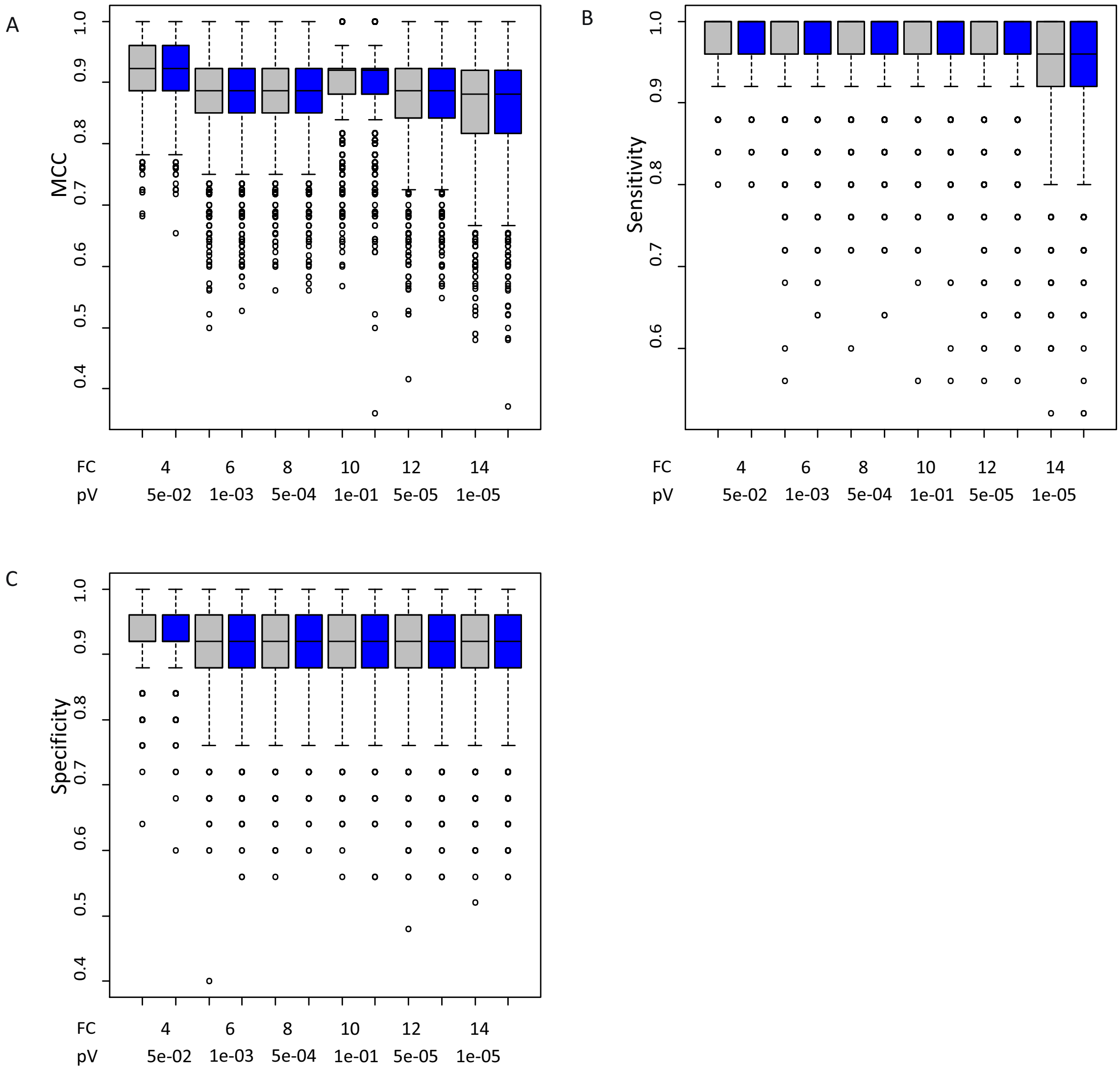


Figure B.0.1.: Prediction performance of small AML
MCC (A), sensitivity (B) and specificity (C) values for 10,000 iterations of classification of the small AML cohort for VS1 (grey) and VS2 (blue) summarized in boxplots for different feature size cutoffs (FC/p-value filter). Boxes indicate the 25th and 75th percentiles, the line within the box marks the median and whiskers above and below the box indicate the 90th and 10th percentiles. Outliers are plotted as dots.

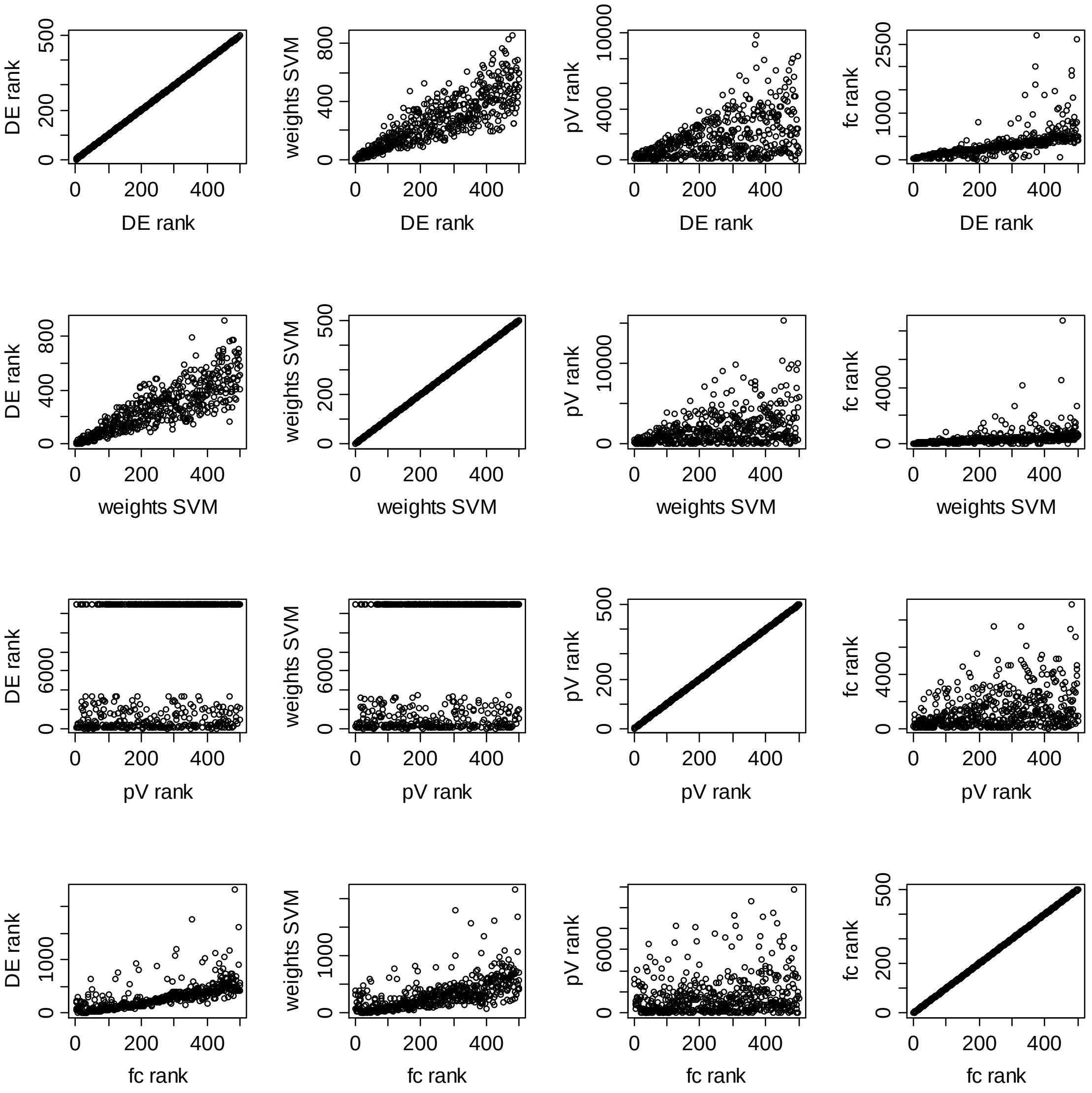


Figure B.0.4.: Comparison of AML feature ranking

All features are ranked using the four different methods described in chapter 3, namely DE selection, weights SVM, p-value and FC ranking. Feature ranking is performed using 10,000 iterations of TSA for the small AML PTS. Shown are the top 500 ranked features for each individual method in comparison to the ranking obtained in the three other methods to show similarity of obtained feature rankings by the different methods. DE ranking and weights SVM ranking are more similar than the other two methods and the p-value resulted in the least similar feature ranking.

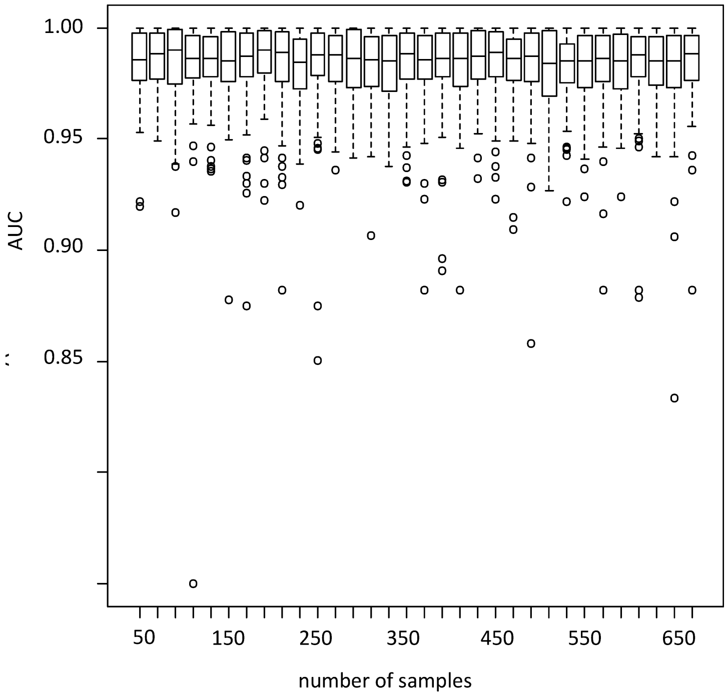


Figure B.0.5.: Adaptive learning of simulated samples based on the method of (*Parrish et al.*, 2009).
Based on the proposed method, 717 AML and 1296 control samples were simulated using information from the small AML PTS. Adaptive learning as described in sec. 3.2.7 was performed using this data set, starting with 50 init TS samples and then subsequentially adding 20 samples. Shown are the boxplots of 100 mean-of-the-mean AUCs for TSA of different sample subsets, comparable to Fig. 5.2.13. Cubic spline fitting for sample size estimation would not be informative in this case.

C. Supplementary methods

C.1. NSCLC GEX data set

Cases and controls

NSCLC cases and hospital based controls were recruited at the University Hospital Cologne and the Lung Clinic Merheim, Cologne, Germany. Healthy blood donors were recruited at the Institute for Transfusion Medicine, University of Cologne. From all individuals PAXgene stabilized blood samples were taken for blood-based gene expression profiling. For all NSCLC cases blood was taken prior chemotherapy. To establish and validate a NSCLC specific classifier 3 independent sets of cases and controls were assembled. The training set (TS) comprised 77 individuals, 35 of those represent NSCLC cases of stage I-IV admitted to the hospital with symptoms of non-small cell lung cancer (coughing, dyspnea, weight loss or reduction in general health state) and 42 were hospital based controls with a comparable comorbidity but no prior history of lung cancer. The validation set 1 (VS1, n=54) likewise contained 28 NSCLC cases of stage I-IV and 26 hospital based controls. Overall, the hospital based controls in TS and VS1 enclosed individuals suffering from advanced chronic obstructive pulmonary disease (COPD) as typically seen in a population of heavily smoking adults (TS: n=7 VS1: n=5). Other diseases such as hypertension (TS: n=17, VS: n=11) or other malignancies (TS: n=10; VS1: n=6) were also observed in the group of hospital based controls. The validation set 2 (VS2, n=102) contained 32 NSCLC cases that had documented stage I NSCLC and were diagnosed mostly during routine chest-rays or due to clinical workup of unspecific symptoms such as reduced general health status. All individuals had an ECOG performance status of 0. In addition VS2 contains 70 healthy blood donors without prior history of lung cancer. The analyses were approved by the local ethics committee and all probands gave informed consent.

Blood collection, cRNA synthesis and array hybridization

Blood (2.5 ml) was drawn into PAXgene vials. After RNA isolation biotin labeled cRNA preparation was performed using the Ambion® Illumina RNA amplification kit (Ambion, UK) or Epicentre TargetAmpTM Kit (Epicentre Biotechnologies, USA) and Biotin-16-UTP (10 mM; Roche Molecular Biochemicals) or Illumina® TotalPrep RNA Amplification Kit (Ambion, UK). Biotin labeled cRNA (1.5 µg) was hybridized to Sentrix® whole genome bead chips WG6 version 2.1 (Illumina, USA) and scanned on the Illumina® BeadStation 500x. For data collection, we used Illumina® BeadStudio 3.1.1.0 software. Data are available at http://www.ncbi.nlm.nih.gov/geo/GSE12771. For RNA quality control the ratio of the OD at wavelengths of 260 nm and 280 nm was calculated for all samples and was between 1.85 and 2. To determine the quality of cRNA, a semi-quantitative RT-PCR amplifying a 5´prime and a 3´prime product of the ß-actin gene was used and demonstrated no sign of degradation with the 5´prime and a 3´prime product being present.

Classification algorithm

Expression values were independently quantile normalized. The classifier for NSCLC was built and optimized based on the TS using a 10 fold CV design. Briefly, TS was divided 10 times into an internal training and an internal validation set in a ratio 9:1. In the internal TS the differentially expressed genes between NSCLC cases and controls were calculated using a T-test. Next 36 different feature lists were extracted from this list of differential expressed genes by 36 times sequentially increasing the cut-off of the p-value (p = 0.00001, p = 0.00002, p = 0.00003 ...-... p = 0.08, p = 0.09, p = 0.1). Subsequently, for each of the resulting 36 feature lists 3 different learning algorithms (SVM, LDA, PAM) were trained on the internal training set and used to calculate the probability score for each case of the respective internal validation set. This approach was repeated 10 times according to the 10 dataset splittings of this 10-fold CV. For each of the 10 CV steps the AUC was calculated for the internal validation set. For each of the 36 cut-offs the mean of the 10 AUCs was calculated. Each of the 10 split data sets was used once as internal validation set. The optimal cut-off p value of the T-statistics and the optimal classification algorithm were selected according to the maximum mean AUC ever reached in all of the three algorithms. We subsequently built a classifier using the respective cut-off p-value of the T-statistics and the selected algorithm in the TS. The classifier was validated in 2 independent validation sets. The AUC was used to measure the quality of the classifier. To test the specificity of the classifier the whole analysis was repeated thousand times using random feature sets of equal size. For visualization of the test score obtained by the SVM algorithm we used the following transformation algorithm: $log_2(score + 1) + 0.1$.

C.2. microRNA profiling

Subject information and blood sample collection

Blood samples from 29 apparently healthy blood donors were collected in two Cell Preparation Tubes with sodium citrate (CPT, Becton Dickinson, Heidelberg, Germany) after written informed consent had been obtained and following approval by the institutional review board. Peripheral blood mononuclear cells (PBMC) were prepared following the manufacturer´s protocol. To evaluate the influence of freezing cells in FCS with 10% (v/v) DMSO and storage in liquid nitrogen on miRNA stability, one portion of the freshly isolated PBMC were lysed in TRIZOL® reagent

(1 ml / 1 x 107 PBMC, Invitrogen, Karlsruhe, Germany) and stored at -80°C until further processing. The other portion of isolated PBMC was resuspended in FCS with 10% (v/v) DMSO (1 ml / 1 x 107 PBMC) and frozen at -80°C. The next day frozen PBMC were cryopreserved in liquid nitrogen for several days to weeks until further processing. In order to analyze the difference between total RNA and less abundant low molecular weight (LMW) RNA we collected blood samples from additional six healthy blood doners.

RNA isolation and miRNA microarray procedure

Total RNA from PBMC stored in TRIZOL® was isolated according to the manufacturer´s protocol. For RNA isolation from PBMC stored in liquid nitrogen, cells were removed from liquid nitrogen and transferred to a 37°C water bath until thawing. The thawed cell suspension was quickly transferred to 40 ml chilled RPMI medium and centrifuged at room temperature at 400 x g for 10 min. Supernatant was removed and PBMC were washed once with 50 ml of room temperature RPMI and centrifuged at room temperature at 400 x g for 10 min. Supernatant was completely removed and cells were subsequently lysed in TRIZOL® reagent. RNA isolation was then performed according to the manufacturer´s protocol. Low molecular weight (LMW) RNA molecules were enriched using Invitrogen's PureLink miRNA Isolation Kit (Invitrogen, Karlsruhe, Germany) according to manufacture's protocol. 50 – 350 ng of LMW RNAs were enriched from 1 – 2 ug of total RNAs, and 10 – 70 ng of the enriched RNAs were used for sample labeling and array hybridization. For the comparison of isolation techniques blood samples from 6 apparently healthy blood donors were collected either in Cell Preparation Tubes with sodium citrate (CPT) or in PAXgene Blood RNA Tubes (PreAnalytiX, Hombrechtikon, Switzerland). PAXgene Blood RNA Tubes were stored at -20°C until further processing. PBMC were prepared from CPT following the manufacturer´s protocol and stored in -80°C. Total RNA from PBMC was isolated using TRIZOL® reagent (invitrogen) according to the manufacturer´s protocol. RNA from PAXgene Blood RNA Tubes was isolated using the PAXgene Blood miRNA Beta Version extraction kit. Total RNA was quantified by UV-spectroscopy at 260 nm. The quality of the isolated RNA samples were determined by measuring the A260 / A280 ratio and the integrity of the ribosomal 28s and 18s bands were determined by agarose-gel electrophoresis. MicroRNA expression profiling was performed using the MicroRNA Profiling Beta-Test Assay Kit for Sentrix Array Matrixes (Illumina, CA, USA). This system provides a highly multiplexed assay and 96-sample Sentrix Array Matrix (SAM) readout. The miRNA microarray assays were generally performed with 500 ng total RNA if not otherwise stated. All steps were performed according to the manufacturer´s protocol. After hybridization signal intensities at each address location were measured using Illumina BeadArray Reader 500x (Illumina, CA, USA). The intensities of the signals correspond to the quantity of the respective miRNA in the original sample.

Statistical and bioinformatics analysis

Raw data extraction of miRNA microarrays was performed with Illumina Beadstudio 3.1.1.0 software using the Beadstudio Gene Expression Analysis Module 3.1.8. Prior analysis data quality assessment was performed and samples with lower overall intensity distributions and decreased number of miRNA transcripts detected as present were excluded from further analysis. For further analysis we used quantile normalization implemented in the Bioconductor affy package. Variable miRNAs were defined by a coefficient of variation (SD/mean) between 0.5 - 10. Determination of present calls was based on the detection p-value assessed by Beadstudio software; a miRNA transcript was called present if the detection p-value was $<$ 0.05. Otherwise the miRNA transcript was called absent. Differentially expressed miRNAs were selected using a fold-change/p-value filter. The Benjamini-Hochberg method was used to adjust the raw p-values to control the false discovery rate. Hierarchical cluster analysis was performed using the hcluster method in R. Before clustering, the data were log2 transformed. Distances of the samples were calculated using Pearson correlation and clusters were formed by taking the average of each cluster.

Validation of miRNA expression results

Quantitative (q) PCR analysis of a selected number of miRNA targets was performed on the six aforementioned blood samples from healthy donors. Three RNA isolation approaches were compared: total RNA, enriched small RNA and PAXgene isolated RNA. Twelve miRNAs were selected (hsa-miR-100, 125a, 125b, 135a, 146a, 150,17-3p, 221, 26a, 31, 93 and 328) and data were produced using those RNA samples that were also analysed in array based miRNA profiling. Absolute association of normalized miRNA array expression intensities (log10) versus the negative cycle treshhold (Ct) value was explored via Spearman's correlation coefficient.

D. R scripts

Algorithmus D.1 Validation algorithm

```
# validation function
###################
# author Andrea Hofmann.
###################
# this function
# 1. builds a SMV classifier on training set of cases and controls already filtered
     on DE genes
# 2. applies this classifier to a validation set
# 3. returns AUC, MCC, Sens + Spec
###################
# input:
# data.case.test + data.control.test = entire training data of cases + controls
# data.case.val + data.control.val = entire validation data of cases + controls
# output:
# AUC, MCC, Sens, Spec
library(e1071)
library(multtest)
library(stats)
library(ROCR)
external <- function(data.case.test,data.con.test,data.case.val,data.con.val) {
SVM <- svm(t(cbind(data.case.test,data.con.test)),as.factor(c(rep("case",ncol(data.
    case.test)),rep("Con",ncol(data.con.test)))),probability=TRUE)
svm.pred <- predict(SVM,t(cbind(data.case.val,data.con.val)),probability=T)
pred.dist <- attr(svm.pred,"probabilities")[,which(colnames(attr(svm.pred,"
    probabilities"))=="Con")]
pred.rocr <- prediction(pred.dist,as.factor(c(rep("case",ncol(data.case.val)),rep("
    Con",ncol(data.con.val)))))
perf.rocr <- performance(pred.rocr,measure="auc")
AUC <- as.numeric(perf.rocr@y.values)
MCC <- performance(pred.rocr,measure="phi")@y.values[[1]][[which((performance(pred.
    rocr,measure="sens")@y.values[[1]]+performance(pred.rocr,measure="spec")@y.
    values[[1]]-1)==max(performance(pred.rocr,measure="sens")@y.values[[1]]+
    performance(pred.rocr,measure="spec")@y.values[[1]]-1))[1]]]
SENS <- performance(pred.rocr,measure="spec")@y.values[[1]][[which((performance(
    pred.rocr,measure="sens")@y.values[[1]]+performance(pred.rocr,measure="spec")@y
    .values[[1]]-1)==max(performance(pred.rocr,measure="sens")@y.values[[1]]+
    performance(pred.rocr,measure="spec")@y.values[[1]]-1))[1]]]
SPEC <- performance(pred.rocr,measure="sens")@y.values[[1]][[which((performance(
    pred.rocr,measure="sens")@y.values[[1]]+performance(pred.rocr,measure="spec")@y
    .values[[1]]-1)==max(performance(pred.rocr,measure="sens")@y.values[[1]]+
    performance(pred.rocr,measure="spec")@y.values[[1]]-1))[1]]]
return(cbind(AUC,MCC,SENS,SPEC)) }
```

Algorithmus D.2 Feature selection algorithms

```
# feature selection function
##################
# author Andrea Hofmann.
##################
# this function
# 1. calculates DE genes based on given p-value, FC and teststatistic for 2 groups
# 2. returns list of DE genes
##################
# input:
# sample.1, sample.2 = data for 2 groups
# fc=2, pval=0.05, teststat="t" = variables with default values
# output:
# list of DE genes
library(multtest)
library(stats)
getDEgenes <- function(sample1,sample2,fc=2,pval=0.05) {
mean_s1 <- apply(sample1,1,mean)
mean_s2 <- apply(sample2,1,mean)
fc_s1_vs_s2 <- getfoldchange(mean_s1,mean_s2,fc)
teststat_s1_vs_s2 <- mt.teststat(cbind(sample1,sample2),c(rep(0,dim(sample1)[2]),
    rep(1,dim(sample2)[2])),test="t")
pvals_s1_vs_s2 <- 2*(1-pnorm(abs(teststat_s1_vs_s2)))
pval_adjust <- p.adjust(pvals_s1_vs_s2, method = "BH")
DEgenes_index <- intersect(which(fc_s1_vs_s2[,2]==1),which(pval_adjust<pval))
return(rownames(sample1[DEgenes_index,])) }
getfoldchange <- function(sample1,sample2,fc) {
t(apply(as.matrix(cbind(sample1,sample2)),1,isfoldchange,fc)) }
isfoldchange <- function(sample,fc) {
change.fc <- function(value)   {
if (value < 1)      {         new_value <- -(1/value)      }
else      {         new_value <- value      }
return(new_value)    }
isdifferent <- 0
fold <- sample[1]/sample[2]
if (is.na(fold))    {      fold<-1    }
if (fold > fc || fold < (1/fc))    {
isdifferent <- 1    }
result <- c(sapply(fold,change.fc),isdifferent)
return(result) }
```

Algorithmus D.3 Prediction algorithm

```
# general function for prediction
##################
# author Andrea Hofmann.
##################
# this function
# 1. requires a data set of case and control samples and an external case/control
    set
# 2. divides the data set into 1 training and 2 validation sets based on given
    proportions with or withour replacement
# 3. calculates differentially expressed genes in the training set
# 3. only goes on if number d.e. genes >=2 (required for classifier)
# 4. builds a classifier in the training set
# 5. applies classifier to 2 validation sets + external set
# 6. reports DE genes + AUC.MCC.Sensitivity.Specificity of prediction of 2
    validation + external sets
##################
# input:
# data.case, data.control = entire dataset of cases + controls
# data.case, data.control = entire dataset of cases + controls
# size.ts.case + size.ts.con = number of cases in training set for cases and
    controls
# size.vs.case + size.vs.con = number of cases in validation set for cases and
    controls
```

```
# ext.vs,classes.ext = external data + corresponding classes
# c_fc + c_pval = fold change + p-value for gene selection
# c_test = test statistic for gene selection. "t" for Welch t-test (unequal
    variances). "wilcoxon" for rank sum Wilcoxon test
# c_classifier = classification algorithm. "SVM". "PAM". "LDA"
# c_kernel = kernel for SVM classification. "linear". "polynomial". radial"."
    sigmoid"
# withoutReplacement = parameter T/F for sample split or bootstrap
###################
# output:
# DE.ts, length(DE.ts), pvals, FC, weights_SVM, AUC_VS1, AUC_VS2, MCC_VS1, MCC_VS2,
     SENS_VS1, SENS_VS2, SPEC_VS1, SPEC_VS2,
# AUC_EXT, MCC_EXT, SENS_EXT, SPEC_EXT
###################
# call:
# prediction.function(parameters)
# permutation:
# permutation.results <- replicate(10000.permutation.function(parameters).simplify=
    FALSE)
library(e1071)
library(multtest)
library(stats)
library(ROCR)
permutation.function <- function( data.case,data.control,size.ts.case,size.ts.con,
    size.vs.case,size.vs.con,c_fc=5,c_pval=0.001,c_test="t",
                                    c_classifier="SVM",c_kernel="linear",ext.vs,
    classes.ext,withoutReplacement=TRUE)
{
if(withoutReplacement){
ds1.case <- sample(1:ncol(data.case),size.ts.case)
ds1.con <- sample(1:ncol(data.control),size.ts.con)
ds2.case <- sample((1:ncol(data.case))[-ds1.case],size.vs.case)
ds2.con <- sample((1:ncol(data.control))[-ds1.con],size.vs.con)
ds3.case <- sample((1:ncol(data.case))[-c(ds1.case,ds2.case)],size.vs.case)
ds3.con <- sample((1:ncol(data.control))[-c(ds1.con,ds2.con)],size.vs.con)    }
else{    random.case <- sample(1:ncol(data.case),ncol(data.case),replace=TRUE)
random.con <- sample(1:ncol(data.control),ncol(data.control),replace=TRUE)
ds1.case <- random.case[1:size.ts.case]
ds1.con <- random.con[1:size.ts.con]
ds2.case <- random.case[(size.ts.case+1):(size.ts.case+size.vs.case)]
ds2.con <- random.con[(size.ts.con+1):(size.ts.con+size.vs.con)]
ds3.case <- random.case[(size.ts.case+size.vs.case+1):ncol(data.case)]
ds3.con <- random.con[(size.ts.con+size.vs.con+1):ncol(data.control)]
  }
ts <- cbind(data.case[,ds1.case],data.control[,ds1.con])
vs1 <- cbind(data.case[,ds2.case],data.control[,ds2.con])
vs2 <- cbind(data.case[,ds3.case],data.control[,ds3.con])
classes.ts <- c(rep("case",length(ds1.case)),rep("Con",length(ds1.con)))
classes.vs1 <- c(rep("case",length(ds2.case)),rep("Con",length(ds2.con)))
classes.vs2 <- c(rep("case",length(ds3.case)),rep("Con",length(ds3.con)))
DE.ts.all <- getDEgenes(ts[,which(classes.ts=="case")],ts[,which(classes.ts=="Con")
    ],fc=c_fc,pval=c_pval,teststat=c_test)
DE.ts <- DE.ts.all[[1]]
if(length(DE.ts)<2)    {
auc.vs1 <- "NA"
mcc.vs1 <- "NA"
auc.vs2 <- "NA"
mcc.vs2 <- "NA"
sens.vs1 <- "NA"
spec.vs1 <- "NA"
sens.vs2 <- "NA"
spec.vs2 <- "NA"
w_SVM <- "NA"
auc.ext <- "NA"
mcc.ext <- "NA"
sens.ext <- "NA"
spec.ext <- "NA"    }
else{
if(c_classifier=="SVM")    {
SVM.ts = svm(t(ts[DE.ts,]),as.factor(classes.ts),kernel=c_kernel,cross=10,
    probability = TRUE)
w_SVM <- t(SVM.ts$coefs) %*% SVM.ts$SV
svm.pred <- predict(SVM.ts,t(vs1[DE.ts,]),probability=T)
pred <- attr(svm.pred,"probabilities")
pred.dist <- pred[,which(colnames(pred)=="Con")]
pred.rocr <- prediction(pred.dist,as.factor(classes.vs1))
perf.rocr <- performance(pred.rocr,measure="auc")
auc.vs1 <- as.numeric(perf.rocr@y.values)
mcc.vs1 <- performance(pred.rocr,measure="phi")@y.values[[1]][which((performance(
    pred.rocr,measure="sens")@y.values[[1]]+performance(pred.rocr,measure="spec")@y
    .values[[1]]-1)==max(performance(pred.rocr,measure="sens")@y.values[[1]]+
    performance(pred.rocr,measure="spec")@y.values[[1]]-1))]
sens.vs1 <- performance(pred.rocr,measure="spec")@y.values[[1]][which((performance(
    pred.rocr,measure="sens")@y.values[[1]]+performance(pred.rocr,measure="spec")@y
    .values[[1]]-1)==max(performance(pred.rocr,measure="sens")@y.values[[1]]+
    performance(pred.rocr,measure="spec")@y.values[[1]]-1))]
spec.vs1 <- performance(pred.rocr,measure="sens")@y.values[[1]][which((performance(
    pred.rocr,measure="sens")@y.values[[1]]+performance(pred.rocr,measure="spec")@y
    .values[[1]]-1)==max(performance(pred.rocr,measure="sens")@y.values[[1]]+
    performance(pred.rocr,measure="spec")@y.values[[1]]-1))]
svm.pred <- predict(SVM.ts,t(vs2[DE.ts,]),probability=T)
pred <- attr(svm.pred,"probabilities")
pred.dist <- pred[,which(colnames(pred)=="Con")]
pred.rocr <- prediction(pred.dist,as.factor(classes.vs2))
perf.rocr <- performance(pred.rocr,measure="auc")
auc.vs2 <- as.numeric(perf.rocr@y.values)
mcc.vs2 <- performance(pred.rocr,measure="phi")@y.values[[1]][which((performance(
    pred.rocr,measure="sens")@y.values[[1]]+performance(pred.rocr,measure="spec")@y
    .values[[1]]-1)==max(performance(pred.rocr,measure="sens")@y.values[[1]]+
    performance(pred.rocr,measure="spec")@y.values[[1]]-1))]
sens.vs2 <- performance(pred.rocr,measure="spec")@y.values[[1]][which((performance(
    pred.rocr,measure="sens")@y.values[[1]]+performance(pred.rocr,measure="spec")@y
    .values[[1]]-1)==max(performance(pred.rocr,measure="sens")@y.values[[1]]+
    performance(pred.rocr,measure="spec")@y.values[[1]]-1))]
spec.vs2 <- performance(pred.rocr,measure="sens")@y.values[[1]][which((performance(
    pred.rocr,measure="sens")@y.values[[1]]+performance(pred.rocr,measure="spec")@y
    .values[[1]]-1)==max(performance(pred.rocr,measure="sens")@y.values[[1]]+
    performance(pred.rocr,measure="spec")@y.values[[1]]-1))]
### independent validation set
svm.pred <- predict(SVM.ts,t(ext.vs[DE.ts,]),probability=T)
pred <- attr(svm.pred,"probabilities")
pred.dist <- pred[,which(colnames(pred)=="Con")]
pred.rocr <- prediction(pred.dist,as.factor(classes.ext))
perf.rocr <- performance(pred.rocr,measure="auc")
auc.ext <- as.numeric(perf.rocr@y.values)
mcc.ext <- performance(pred.rocr,measure="phi")@y.values[[1]][which((performance(
    pred.rocr,measure="sens")@y.values[[1]]+performance(pred.rocr,measure="spec")@y
    .values[[1]]-1)==max(performance(pred.rocr,measure="sens")@y.values[[1]]+
    performance(pred.rocr,measure="spec")@y.values[[1]]-1))]
sens.ext <- performance(pred.rocr,measure="spec")@y.values[[1]][which((performance(
    pred.rocr,measure="sens")@y.values[[1]]+performance(pred.rocr,measure="spec")@y
    .values[[1]]-1)==max(performance(pred.rocr,measure="sens")@y.values[[1]]+
    performance(pred.rocr,measure="spec")@y.values[[1]]-1))]
spec.ext <- performance(pred.rocr,measure="sens")@y.values[[1]][which((performance(
    pred.rocr,measure="sens")@y.values[[1]]+performance(pred.rocr,measure="spec")@y
    .values[[1]]-1)==max(performance(pred.rocr,measure="sens")@y.values[[1]]+
    performance(pred.rocr,measure="spec")@y.values[[1]]-1))]    }    }
return(list(probes = DE.ts,number_probes = length(DE.ts),pvals = DE.ts.all[[2]], FC
     = DE.ts.all[[3]], weights_SVM = w_SVM, AUC_VS1 = auc.vs1,AUC_VS2 = auc.vs2,
    MCC_VS1 = mcc.vs1,MCC_VS2 = mcc.vs2,SENS_VS1 = sens.vs1, SENS_VS2 = sens.vs2,
    SPEC_VS1 = spec.vs1,SPEC_VS2 = spec.vs2, AUC_EXT = auc.ext,MCC_EXT = mcc.ext,
    SENS_EXT = sens.ext, SPEC_EXT = spec.ext)) }
```

Bibliography

Abeel, T., T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys (2010), Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, *Bioinformatics*, *26*(3), 392–398.

Alizadeh, A. A., et al. (2000), Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, *403*, 503–511.

Allison, D. B., X. Cui, G. P. Page, and M. Sabripour (2006), Microarray data analysis: from disarray to consolidation and consensus, *Nature Genetics*, *7*, 55–65.

American Cancer Society (2011), *Cancer Facts and Figures 2011*, Atlanta, GA, USA.

American Cancer Society (2012), http://www.cancer.org/.

Ashburner, M., et al. (2000), Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat. Genet.*, *25*(1), 25–29.

Avet-Loiseau, H., et al. (2009), Prognostic significance of copy-number alterations in multiple myeloma, *J. Clin. Oncol.*, *27*(27), 4585–4590.

Bach, P. B., G. A. Silvestri, M. Hanger, and J. R. Jett (2007), Screening for lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition), *Chest*, *132*, 69S–77S.

Bacher, U., A. Kohlmann, C. Haferlach, and T. Haferlach (2009a), Gene expression profiling in acute myeloid leukaemia (AML), *Best Pract Res Clin Haematol*, *22*(2), 169–180.

Bacher, U., A. Kohlmann, and T. Haferlach (2009b), Current status of gene expression profiling in the diagnosis and management of acute leukaemia, *Br. J. Haematol.*, *145*(5), 555–568.

Baird, A. E. (2006), The blood option: transcriptional profiling in clinical trials, *Pharmacogenomics*, *7*, 141–144.

Barbosa-Morais, N. L., M. J. Dunning, S. A. Samarajiwa, J. F. Darot, M. E. Ritchie, A. G. Lynch, and S. Tavare (2010), A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data, *Nucleic Acids Res.*, *38*(3), e17.

Barnes, M. G., et al. (2009), Subtype-specific peripheral blood gene expression profiles in recent-onset juvenile idiopathic arthritis, *Arthritis Rheum.*, *60*(7), 2102–2112.

Barrett, T., et al. (2011), NCBI GEO: archive for functional genomics data sets–10 years on, *Nucleic Acids Res.*, *39*(Database issue), D1005–1010.

Barry, C. E., H. I. Boshoff, V. Dartois, T. Dick, S. Ehrt, J. Flynn, D. Schnappinger, R. J. Wilkinson, and D. Young (2009), The spectrum of latent tuberculosis: rethinking the biology and intervention strategies, *Nat. Rev. Microbiol.*, *7*(12), 845–855.

Bartel, D. P. (2004), MicroRNAs: genomics, biogenesis, mechanism, and function, *Cell*, *116*(2), 281–297.

Baty, F., M. Facompré, J. Wiegand, J. Schwager, and M. H. Brutsche (2006), Analysis with respect to instrumental variables for the exploration of microarray data structures, *BMC Bioinformatics*, *7*, 422–422.

Beer, D. G., et al. (2002), Gene-expression profiles predict survival of patients with lung adenocarcinoma, *Nat. Med.*, *8*, 816–824.

Benjamini, Y., and Y. Hochberg (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society Series B*, *57*, 289ï¿œ300.

Berrar, D., and P. Flach (2012), Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them), *Brief. Bioinformatics*, *13*(1), 83–97.

Berrar, D., I. Bradbury, and W. Dubitzky (2006), Avoiding model selection bias in small-sample genomic datasets, *Bioinformatics*, *22*(10), 1245–1250.

Berry, M. P. R., et al. (2010), An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis, *Nature*, *466*, 973–977.

Bittner, M., et al. (2000), Molecular classification of cutaneous malignant melanoma by gene expression profiling, *Nature*, *406*, 536–540.

Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed (2002), A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics*, *19*(2), 185–193.

Bonferroni, C. E. (1936), Teoria statistica delle classi e calcolo delle probabilita, *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, *8*, 3–62.

Borovecki, F., et al. (2005), Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease, *Proc Natl Acad Sci U S A*, *102*, 11,023–11,028.

Boulesteix, A. L., and M. Slawski (2009), Stability and aggregation of ranked gene lists, *Brief. Bioinformatics*, *10*(5), 556–568.

Bradley, A. P. (1997), The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms, *Pattern Recognition*, *30*, 1145–1159.

Braga-Neto, U. M., and E. R. Dougherty (2004), Is cross-validation valid for small-sample microarray classification?, *Bioinformatics*, *20*(3), 374–380.

Brazma, A., et al. (2001), Minimum information about a microarray experiment (MIAME)-toward standards for microarray data, *Nat. Genet.*, *29*, 365–371.

Breitling, R., P. Armengaud, A. Amtmann, and P. Herzyk (2004), Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments, *FEBS Letters*, *573*, 83–92.

Brenner, H., A. Gondos, and D. Pulte (2008), Recent major improvement in long-term survival of younger patients with multiple myeloma, *Blood*, *111*(5), 2521–2526.

Brenner, S., F. Jacob, and M. Meselson (1961), An unstable intermediate carrying information from genes to ribosomes for protein synthesis, *Nature*, *190*, 576–581.

Broyl, A., et al. (2010), Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients, *Blood*, *116*(14), 2543–2553.

Bueno-de Mesquita, J. M., et al. (2007), Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER), *Lancet Oncol.*, *8*, 1079–1087.

Bullinger, L., K. Dohner, E. Bair, S. Frohling, R. F. Schlenk, R. Tibshirani, H. Dohner, and J. R. Pollack (2004), Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia, *N. Engl. J. Med.*, *350*, 1605–1616.

Burczynski, M. E., and A. J. Dorner (2006), Transcriptional profiling of peripheral blood cells in clinical pharmacogenomic studies, *Pharmacogenomics*, *7*(2), 187–202.

Burczynski, M. E., et al. (2005), Transcriptional profiles in peripheral blood mononuclear cells prognostic of clinical outcomes in patients with advanced renal cell carcinoma, *Clin. Cancer Res.*, *11*, 1181–1189.

Calin, G. A., and C. M. Croce (2006), MicroRNA signatures in human cancers, *Nat. Rev. Cancer*, *6*(11), 857–866.

Calin, G. A., et al. (2004), MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias, *Proc. Natl. Acad. Sci. U.S.A.*, *101*(32), 11,755–11,760.

Calin, G. A., et al. (2005), A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia, *N. Engl. J. Med.*, *353*(17), 1793–1801.

Canales, R. D., et al. (2006), Evaluation of DNA microarray results with quantitative gene expression platforms, *Nat. Biotechnol.*, *24*(9), 1115–1122.

Cancer Research UK (2012), http://www.cancerresearchuk.org/.

Cardoso, F., L. Van't Veer, E. Rutgers, S. Loi, S. Mook, and M. J. Piccart-Gebhart (2008), Clinical application of the 70-gene profile: the MINDACT trial, *J. Clin. Oncol.*, *26*, 729–735.

Chang, C., et al. (2011), Maximizing biomarker discovery by minimizing gene signatures, *BMC Genomics*, *12 Suppl 5*, S6.

Chang, H. H., and M. F. Ramoni (2009), Transcriptional network classifiers, *BMC Bioinformatics*, *10 Suppl 9*, S1.

Chaussabel, D., V. Pascual, and J. Banchereau (2010), Assessing the human immune system through blood transcriptomics, *BMC Biol.*, *8*, 84.

Chen, J., J. Lozach, E. W. Garcia, B. Barnes, S. Luo, I. Mikoulitch, L. Zhou, G. Schroth, and J. B. Fan (2008), Highly sensitive and specific microRNA expression profiling using BeadArray technology, *Nucleic Acids Res.*, *36*(14), e87.

Chen, J., O. Odenike, and J. D. Rowley (2010), Leukaemogenesis: more than mutant genes, *Nat. Rev. Cancer*, *10*(1), 23–36.

Chen, M., L. Shi, R. Kelly, R. Perkins, H. Fang, and W. Tong (2011), Selecting a single model or combining multiple models for microarray-based classifier development?–a comparative analysis based on large and diverse datasets generated from the MAQC-II project, *BMC Bioinformatics*, *12 Suppl 10*, S3.

Cheson, B. D., et al. (2003), Revised recommendations of the International Working Group for Diagnosis, Standardization of Response Criteria, Treatment Outcomes, and Reporting Standards for Therapeutic Trials in Acute Myeloid Leukemia, *J. Clin. Oncol.*, *21*(24), 4642–4649.

Chisholm, E., U. Bapat, C. Chisholm, G. Alusi, and G. Vassaux (2007), Gene therapy in head and neck cancer: a review, *Postgrad Med J*, *83*, 731–737.

Chuang, H. Y., E. Lee, Y. T. Liu, D. Lee, and T. Ideker (2007), Network-based classification of breast cancer metastasis, *Mol. Syst. Biol.*, *3*, 140.

Chung, W. H., et al. (2008), Granulysin is a key mediator for disseminated keratinocyte death in Stevens-Johnson syndrome and toxic epidermal necrolysis, *Nat. Med.*, *14*(12), 1343–1350.

Connolly, P. H., V. J. Caiozzo, F. Zaldivar, D. Nemet, J. Larson, S.-P. Hung, J. D. Heck, G. W. Hatfield, and D. M. Cooper (2004), Effects of exercise on gene expression in human peripheral blood mononuclear cells, *J Appl Physiol*, *97*, 1461–1469.

Crick, F. H., L. Barnett, S. Brenner, and R. J. Watts-Tobin (1961), General nature of the genetic code for proteins, *Nature*, *192*, 1227–1232.

Critchley-Thorne, R. J., N. Yan, S. Nacu, J. Weber, S. P. Holmes, and P. P. Lee (2007), Down-regulation of the interferon signaling pathway in T lymphocytes from patients with metastatic melanoma, *PLoS Med.*, *4*, e176.

Cronin, M., and J. S. Ross (2011), Comprehensive next-generation cancer genome sequencing in the era of targeted therapy and personalized oncology, *Biomark Med*, *5*(3), 293–305.

Cun, Y., and H. Frohlich (2012), Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions, *BMC Bioinformatics*, *13*(1), 69.

Davis, C. A., F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Kuffner, and R. Zimmer (2006), Reliable gene signatures for microarray classification: assessment of stability and performance, *Bioinformatics*, *22*(19), 2356–2363.

De Jager, P. L., et al. (2009), Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci, *Nat. Genet.*, *41*(7), 776–782.

de Jonge, H. J., et al. (2010), High VEGFC expression is associated with unique gene expression profiles and predicts adverse prognosis in pediatric and adult acute myeloid leukemia, *Blood*, *116*(10), 1747–1754.

Debey, S., U. Schoenbeck, M. Hellmich, B. S. Gathof, R. Pillai, T. Zander, and J. L. Schultze (2004), Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on overall expression and the role of different cell types, *Pharmacogenomics J.*, *4*, 193–207.

Debey, S., T. Zander, B. Brors, A. Popov, R. Eils, and J. L. Schultze (2006), A highly standardized, robust, and cost-effective method for genome-wide transcriptome analysis of peripheral blood applicable to large-scale clinical trials, *Genomics*, *87*, 653–664.

Debey-Pascher, S., A. Hofmann, F. Kreusch, G. Schuler, B. Schuler-Thurner, J. L. Schultze, and A. Staratschek-Jox (2011), RNA-stabilized whole blood samples but not peripheral blood mononuclear cells can be stored for prolonged time periods prior to transcriptome analysis, *J Mol Diagn*, *13*, 452–460.

Ding, L., M. C. Wendl, D. C. Koboldt, and E. R. Mardis (2010), Analysis of next-generation genomic data in cancer: accomplishments and challenges, *Hum. Mol. Genet.*, *19*(R2), R188–196.

Dougherty, E. R., and M. Brun (2006), On the number of close-to-optimal feature sets, *Cancer Inform*, *2*, 189–196.

Du, P., G. Feng, J. Flatow, J. Song, M. Holko, W. A. Kibbe, and S. M. Lin (2009), From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations, *Bioinformatics*, *25*(12), i63–68.

Duan, K. B., J. C. Rajapakse, H. Wang, and F. Azuaje (2005), Multiple SVM-RFE for gene selection in cancer classification with expression data, *IEEE Trans Nanobioscience*, *4*(3), 228–234.

Dudoit, S., J. Fridlyand, and T. Speed (2002a), Comparison of discrimination methods for the classification of tumors using gene expression data., *J. Am. Stat. Assoc.*, *97*, 77–87.

Dudoit, S., Y. H. Yang, M. J. Callow, and T. P. Speed (2002b), Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica*, *12*, 111–139.

Dunning, M. J., M. E. Ritchie, N. L. Barbosa-Morais, S. Tavare, and A. G. Lynch (2008), Spike-in validation of an Illumina-specific variance-stabilizing transformation, *BMC Res Notes*, *1*, 18.

Dupuy, A., and R. M. Simon (2007), Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting, *J. Natl. Cancer Inst.*, *99*, 147–157.

Eden, E., R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini (2009), GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists, *BMC Bioinformatics*, *10*, 48.

Edgar, R., M. Domrachev, and A. E. Lash (2002), Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Res.*, *30*(1), 207–210.

Efron, B., and R. Tibshirani (1993), *An Introduction to the Bootstrap*, Chapman and Hall, London.

Efron, B., and R. Tibshirani (1997), Improvements on Cross-Validation: The .632+ Bootstrap Method, *J. Amer. Statist. Assoc.*, *92*, 548–560.

Efron, B., and R. Tibshirani (2002), Empirical bayes methods and false discovery rates for microarrays, *Genet. Epidemiol.*, *23*, 70–86.

Eggle, D., S. Debey-Pascher, M. Beyer, and J. L. Schultze (2009), The development of a comparison approach for Illumina bead chips unravels unexpected challenges applying newest generation microarrays, *BMC Bioinformatics*, *10*, 186.

Ein-Dor, L., I. Kela, G. Getz, D. Givol, and E. Domany (2005), Outcome signature genes in breast cancer: is there a unique set?, *Bioinformatics*, *21*, 171–178.

Ein-Dor, L., O. Zuk, and E. Domany (2006), Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer, *Proc. Natl. Acad. Sci. U.S.A.*, *103*(15), 5923–5928.

Elashoff, M. R., R. Nuttall, P. Beineke, M. H. Doctolero, M. Dickson, A. M. Johnson, S. E. Daniels, S. Rosenberg, and J. A. Wingrove (2012), Identification of factors contributing to variability in a blood-based gene expression test, *PLoS ONE*, *7*(7), e40,068.

Fall, N., et al. (2007), Gene expression profiling of peripheral blood from patients with untreated new-onset systemic juvenile idiopathic arthritis reveals molecular heterogeneity that may predict macrophage activation syndrome, *Arthritis Rheum.*, *56*(11), 3793–3804.

Fan, C., D. S. Oh, L. Wessels, B. Weigelt, D. S. Nuyten, A. B. Nobel, L. J. van't Veer, and C. M. Perou (2006), Concordance among gene-expression-based predictors for breast cancer, *N. Engl. J. Med.*, *355*, 560–569.

Fan, X., L. Shi, H. Fang, Y. Cheng, R. Perkins, and W. Tong (2010a), DNA microarrays are predictive of cancer prognosis: a re-evaluation, *Clin. Cancer Res.*, *16*, 629–636.

Fan, X., et al. (2010b), Consistency of predictive signature genes and classifiers generated using different microarray platforms, *Pharmacogenomics J.*, *10*(4), 247–257.

Faraggi, D., and B. Reiser (2002), Estimation of the area under the ROC curve, *Stat Med*, *21*, 3093–3106.

FDA (2005), Guidance for industry: Pharmacogenomics data submissions., *Food and Drug Administration, U.S. Department of Health and Human Services.*

Fisher, R. A. (1935), *The Design of Experiments*, Hafner Press, New York.

Frank, M. B., et al. (2009), Disease-associated pathophysiologic structures in pediatric rheumatic diseases show characteristics of scale-free networks seen in physiologic systems: implications for pathogenesis and treatment, *BMC Med Genomics*, *2*, 9.

Freidlin, B., and R. Simon (2005), Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients, *Clin. Cancer Res.*, *11*(21), 7872–7878.

Furey, T. S., N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler (2000), Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, *16*, 906–914.

Gaarz, A., et al. (2010), Bead array-based microrna expression profiling of peripheral blood and the impact of different RNA isolation approaches, *J Mol Diagn*, *12*, 335–344.

Ge, Y., S. Dudoit, and T. Speed (2003), Resampling-based multiple testing for microarray data analysis, *Test*, *12*, 1–77.

Gentleman, R. C., et al. (2004), Bioconductor: open software development for computational biology and bioinformatics, *Genome Biol*, *5*, R80–R80.

Glas, A. M., et al. (2006), Converting a breast cancer microarray signature into a high-throughput diagnostic test, *BMC Genomics*, *7*, 278.

Golub, G., . Heath, and G. Wahba (1979), Generalized cross-validation as a method for choosing a good ridge parameter., *Technometrics*, *21*, 215–224.

Golub, T., D. K. Slonim, and Tamayo, P. et al. (1999), Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, *286*(5349), 531–537.

Gow, J. W., S. Hagan, P. Herzyk, C. Cannon, P. O. Behan, and A. Chaudhuri (2009), A gene signature for post-infectious chronic fatigue syndrome, *BMC Med Genomics*, *2*, 38–38.

Guo, L., et al. (2006), Rat toxicogenomic study reveals analytical consistency across microarray platforms, *Nat. Biotechnol.*, *24*(9), 1162–1169.

Gutiérrez, N. C., et al. (2005), Gene expression profile reveals deregulation of genes with relevant functions in the different subclasses of acute myeloid leukemia, *Leukemia*, *19*, 402–409.

Gutteridge, A., and J. M. Thornton (2005), Understanding nature's catalytic toolkit, *Trends Biochem. Sci.*, *30*(11), 622–629.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik (2002), Gene selection for cancer classification using support vector machines, *Machine learning*, *46*, 389–422, original SVM-RFE paper.

Haferlach, T., A. Kohlmann, S. Schnittger, M. Dugas, W. Hiddemann, W. Kern, and C. Schoch (2005), Global approach to the diagnosis of leukemia using gene expression profiling, *Blood*, *106*, 1189–1198.

Hanczar, B., J. Hua, C. Sima, J. Weinstein, M. Bittner, and E. R. Dougherty (2010), Small-sample precision of ROC-related estimates, *Bioinformatics*, *26*(6), 822–830.

Hand, D. J., and R. J. Till (2001), A simple generalization of the area under the ROC curve to multiple class classification problems, *Machine Learning*, *45*(2).

Hastie, T., R. Tibshirani, and J. Friedman (2001), *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, Springer, NY, USA.

Hastie, T. J., and R. Tibshirani (1990), *Generalized Additive Models*, Chapman and Hall, London.

He, H. (2009), Learning from Imbalanced Data, *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

He, Z., and W. Yu (2010), Stable feature selection for biomarker discovery, *Comput Biol Chem*, *34*(4), 215–225.

Henschke, C. I., D. F. Yankelevitz, D. M. Libby, M. W. Pasmantier, J. P. Smith, and O. S. Miettinen (2006), Survival of patients with stage I lung cancer detected on CT screening, *N. Engl. J. Med.*, *355*, 1763–1771.

Henseler, M. (2009), *Using whole genome wide gene expression profiling for the establishment of disease specific gene signatures in peripheral blood*, Master Thesis, RheinAhrCampus Remagen, University of Applied Sciences Koblenz.

Hilden, J. (2005), What properties should an overall measure of test performance possess?, *Clin. Chem.*, *51*(2), 471–472.

Holm, S. (1979), A Simple Sequentially Rejective Bonferroni Test Procedure, *Scandinavian Journal of Statistics*, *6*, 65–70.

Hood, L., and D. Galas (2003), The digital code of DNA, *Nature*, *421*, 444–448.

Hua, J., Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty (2005), Optimal number of features as a function of sample size for various classification rules, *Bioinformatics*, *21*(8), 1509–1515.

Huang, J., et al. (2010), Genomic indicators in the blood predict drug-induced liver injury, *Pharmacogenomics J.*, *10*(4), 267–277.

Hubbell, E., W. M. Liu, and R. Mei (2002), Robust estimators for expression analysis, *Bioinformatics*, *18*(12), 1585–1592.

Huber, W., A. von Heydebreck, H. Sï¿œltmann, A. Poustka, and M. Vingron (2002), Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics*, *18 Suppl.1*, 96–104.

Ihaka, R., and R. Gentlemen (1996), R: a language for data analysis and graphics., *J. Comput. Graph. Stat.*, *5*, 299–314.

Ioannidis, J. P. (2005), Microarrays and molecular research: noise discovery?, *Lancet*, *365*, 454–455.

Ioannidis, J. P. (2007), Is molecular profiling ready for use in clinical decision making?, *Oncologist*, *12*(3), 301–311.

Iorio, M. V., et al. (2005), MicroRNA gene expression deregulation in human breast cancer, *Cancer Res.*, *65*(16), 7065–7070.

Irizarry, R. A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed (2003), Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Res.*, *31*, e15.

Jemal, A., R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M. J. Thun (2008), Cancer statistics, 2008, *CA Cancer J Clin*, *58*, 71–96.

Jemal, A., F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman (2011), Global cancer statistics, *CA Cancer J Clin*, *61*(2), 69–90.

Jiang, W., B. Freidlin, and R. Simon (2007), Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect, *J. Natl. Cancer Inst.*, *99*(13), 1036–1043.

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

Keller, A., P. Leidinger, A. Borries, A. Wendschlag, F. Wucherpfennig, M. Scheffler, H. Huwer, H. P. Lenhof, and E. Meese (2009a), miRNAs in lung cancer - studying complex fingerprints in patient's blood cells by microarray experiments, *BMC Cancer*, *9*, 353.

Keller, A., P. Leidinger, J. Lange, A. Borries, H. Schroers, M. Scheffler, H. P. Lenhof, K. Ruprecht, and E. Meese (2009b), Multiple sclerosis: microRNA expression profiles accurately differentiate patients with relapsing-remitting disease from healthy controls, *PLoS ONE*, *4*(10), e7440.

Khan, J., et al. (2001), Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nat. Med.*, *7*, 673–679.

Kihara, C., et al. (2001), Prediction of sensitivity of esophageal tumors to adjuvant chemotherapy by cDNA microarray analysis of gene-expression profiles, *Cancer Res.*, *61*, 6474–6479.

Kostka, D., and R. Spang (2008), Microarray based diagnosis profits from better documentation of gene expression signatures, *PLoS Comput. Biol.*, *4*(2), e22.

Kuhn, K., et al. (2004), A novel, high-performance random array platform for quantitative gene expression profiling, *Genome Res.*, *14*(11), 2347–2356.

Kuncheva, L. I. (2007), A stability index for feature selection, *AIAP'07: Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference*, pp. 390–395.

Kuo, W. P., et al. (2006), A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies, *Nat. Biotechnol.*, *24*(7), 832–840.

Lawn, S. D., and A. I. Zumla (2011), Tuberculosis, *Lancet*, *378*(9785), 57–72.

Lawrie, C. H., et al. (2007), MicroRNA expression distinguishes between germinal center B cell-like and activated B cell-like subtypes of diffuse large B cell lymphoma, *Int. J. Cancer*, *121*(5), 1156–1161.

Lee, Y., and C. K. Lee (2003), Classification of multiple cancer types by multicategory support vector machines using gene expression data, *Bioinformatics*, *19*(9), 1132–1139.

Lehmann, E. L., and J. P. Romano (2005), *Testing Statistical Hypotheses*, 3 ed., Springer, New York.

Li, T., C. Zhang, and M. Ogihara (2004), A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, *Bioinformatics*, *20*(15), 2429–2437.

Li, X., S. Peng, J. Chen, B. Lu, H. Zhang, and M. Lai (2012), SVM-T-RFE: a novel gene selection algorithm for identifying metastasis-related genes in colorectal cancer using gene expression profiles, *Biochem. Biophys. Res. Commun.*, *419*(2), 148–153.

Li, Z., et al. (2009), Gene expression-based classification and regulatory networks of pediatric acute lymphoblastic leukemia, *Blood*, *114*(20), 4486–4493.

Liew, C. C., J. Ma, H. C. Tang, R. Zheng, and A. A. Dempsey (2006), The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool, *J. Lab. Clin. Med.*, *147*, 126–132.

Lin, W. J., H. M. Hsueh, and J. J. Chen (2010), Power and sample size estimation in microarray studies, *BMC Bioinformatics*, *11*, 48.

Liu, E. T., and K. R. Karuturi (2004), Microarrays and clinical investigations, *N. Engl. J. Med.*, *350*, 1595–1597.

Lockhart, D. J., et al. (1996), Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nat. Biotechnol.*, *14*(13), 1675–1680.

Lu, J., et al. (2005), MicroRNA expression profiles classify human cancers, *Nature*, *435*(7043), 834–838.

Luo, J., et al. (2010), A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data, *Pharmacogenomics J.*, *10*(4), 278–291.

Ma, S., and M. R. Kosorok (2010), Detection of gene pathways with predictive power for breast cancer prognosis, *BMC Bioinformatics*, *11*, 1.

Mao, Y., X. Zhou, D. Pi, Y. Sun, and S. T. Wong (2005), Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection, *J. Biomed. Biotechnol.*, *2005*(2), 160–171.

Martinez-Llordella, M., et al. (2007), Multiparameter immune profiling of operational tolerance in liver transplantation, *Am. J. Transplant.*, *7*(2), 309–319.

Mattie, M. D., et al. (2006), Optimized high-throughput microRNA expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies, *Mol. Cancer*, *5*, 24.

McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition.*, Wiley, New York USA.

McNerney, R., et al. (2012), Tuberculosis diagnostics and biomarkers: needs, challenges, recent advances, and opportunities, *J. Infect. Dis.*, *205 Suppl 2*, S147–158.

Metz, C. E. (1978), Basic principles of ROC analysis., *Sem Nuc Med*, *8*, 283–298.

Metzeler, K. H., et al. (2008), An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia, *Blood*, *112*, 4193–4201.

Michiels, S., S. Koscielny, and C. Hill (2005), Prediction of cancer outcome with microarrays: a multiple random validation strategy, *Lancet*, *365*, 488–492.

Miesner, M., et al. (2010), Multilineage dysplasia (MLD) in acute myeloid leukemia (AML) correlates with MDS-related cytogenetic abnormalities and a prior history of MDS or MDS/MPN but has no independent prognostic relevance: a comparison of 408 cases classified as "AML not otherwise specified" (AML-NOS) or "AML with myelodysplasia-related changes" (AML-MRC), *Blood*, *116*(15), 2742–2751.

Miller, R. M., et al. (2004), Dysregulation of gene expression in the 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine-lesioned mouse substantia nigra, *J. Neurosci.*, *24*, 7445–7454.

Mills, K. I., et al. (2009), Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome, *Blood*, *114*(5), 1063–1072.

Molinaro, A. M., R. Simon, and R. M. Pfeiffer (2005), Prediction error estimation: a comparison of resampling methods, *Bioinformatics*, *21*(15), 3301–3307.

Mullighan, C. G., et al. (2009), Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia, *N Engl J Med*, *360*, 470–480.

Mundra, P. A., and J. C. Rajapakse (2010), SVM-RFE with MRMR filter for gene selection, *IEEE Trans Nanobioscience*, *9*(1), 31–37.

Munshi, N. C., and H. Avet-Loiseau (2011), Genomics in multiple myeloma, *Clin. Cancer Res.*, *17*(6), 1234–1242.

National Cancer Institute (2012), http://www.cancer.gov/.

National Institute of Allergy and Infectious Diseases (2012), http://www.niaid.nih.gov/volunteer/hivlongterm/.

Nelson, P. T., W. X. Wang, B. R. Wilfred, and G. Tang (2008), Technical variables in high-throughput miRNA expression profiling: much work remains to be done, *Biochim. Biophys. Acta*, *1779*(11), 758–765.

Newton, R., A. Deonarine, and L. Wernisch (2011), Creating web applications for spatial epidemiological analysis and mapping in R using Rwui, *Source Code Biol Med*, *6*(1), 6.

Nguyen, D. V., and D. M. Rocke (2002), Multi-class cancer classification via partial least squares with gene expression profiles, *Bioinformatics*, *18*(9), 1216–1226.

Nirenberg, M. W., and J. H. Matthaei (1961), The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides, *Proc. Natl. Acad. Sci. U.S.A.*, *47*, 1588–1602.

Ntzani, E. E., and J. P. Ioannidis (2003), Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment, *Lancet*, *362*, 1439–1444.

Oberthuer, A., et al. (2010), Comparison of performance of one-color and two-color gene-expression analyses in predicting clinical endpoints of neuroblastoma patients, *Pharmacogenomics J.*, *10*(4), 258–266.

Obuchowski, N. A. (2005), ROC analysis, *AJR Am J Roentgenol*, *184*(2), 364–372.

Osborne, J. D., J. Flatow, M. Holko, S. M. Lin, W. A. Kibbe, L. J. Zhu, M. I. Danila, G. Feng, and R. L. Chisholm (2009), Annotating the human genome with Disease Ontology, *BMC Genomics*, *10 Suppl 1*, S6.

Osman, I., et al. (2006), Novel blood biomarkers of human urinary bladder cancer, *Clin. Cancer Res.*, *12*, 3374–3380.

Palumbo, A., and K. Anderson (2011), Multiple myeloma, *N. Engl. J. Med.*, *364*(11), 1046–1060.

Papapanou, P. N., et al. (2007), Periodontal therapy alters gene expression of peripheral blood monocytes, *J. Clin. Periodontol.*, *34*(9), 736–747.

Parrish, R. S., H. J. Spencer, and P. Xu (2009), Distribution modeling and simulation of gene expression data, *Computational Statistics and Data Analysis*, *53*(5), 1650–1660.

Parry, R. M., et al. (2010), k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction, *Pharmacogenomics J.*, *10*(4), 292–309.

Pomeroy, S. L., et al. (2002), Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature*, *415*(6870), 436–442.

Poropatich, K., and D. J. Sullivan (2011), Human immunodeficiency virus type 1 long-term non-progressors: the viral, genetic and immunological basis for disease non-progression, *J. Gen. Virol.*, *92*(Pt 2), 247–268.

Quackenbush, J. (2006a), Computational approaches to analysis of DNA microarray data, *Methods Inf Med*, *45 (Suppl. 1)*, 91–103.

Quackenbush, J. (2006b), Microarray analysis and tumor classification, *The New England Journal of Medicine*, *354;23*, 2463–2472.

Quackenbush, J. (2009), Data reporting standards: making the things we use better, *Genome Med*, *1*(11), 111.

Radom-Aizik, S., F. Zaldivar, S. Y. Leu, and D. M. Cooper (2009), A brief bout of exercise alters gene expression and distinct gene pathways in peripheral blood mononuclear cells of early- and late-pubertal females, *J. Appl. Physiol.*, *107*(1), 168–175.

Ramalho-Santos, M., S. Yoon, Y. Matsuzaki, R. C. Mulligan, and D. A. Melton (2002), : transcriptional profiling of embryonic and adult stem cells, *Science*, *298*, 597–600.

Ramaswamy, S., et al. (2001), Multiclass cancer diagnosis using tumor gene expression signatures, *Proc Natl Acad Sci U S A*, *98*, 15,149–15,154.

Ramilo, O., et al. (2007), Gene expression patterns in blood leukocytes discriminate patients with acute infections, *Blood*, *109*, 2066–2077.

Ransohoff, D. F. (2007), How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design, *J Clin Epidemiol*, *60*(12), 1205–1219.

Rosenwald, A., et al. (2002), The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma, *N. Engl. J. Med.*, *346*, 1937–1947.

Ross, M. E., et al. (2004), Gene expression profiling of pediatric acute myelogenous leukemia, *Blood*, *104*, 3679–3687.

Sandler, A., R. Gray, M. C. Perry, J. Brahmer, J. H. Schiller, A. Dowlati, R. Lilenbaum, and D. H. Johnson (2006), Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer, *N. Engl. J. Med.*, *355*, 2542–2550.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown (1995), Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, *270*, 467–470.

Scherzer, C. R., et al. (2007), Molecular markers of early Parkinson's disease based on gene expression in blood, *Proc Natl Acad Sci U S A*, *104*, 955–960.

Schuster, S. C. (2008), Next-generation sequencing transforms today's biology, *Nat. Methods*, *5*(1), 16–18.

Shapiro, D. E. (1999), The interpretation of diagnostic tests, *Stat Methods Med Res*, *8*, 113–134.

Sharma, P., et al. (2005), Early detection of breast cancer based on gene-expression patterns in peripheral blood cells, *Breast Cancer Res.*, *7*, R634–644.

Shaughnessy, J. D., et al. (2007), A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1, *Blood*, *109*(6), 2276–2284.

Shendure, J., and H. Ji (2008), Next-generation DNA sequencing, *Nat. Biotechnol.*, *26*(10), 1135–1145.

Sheppard, H. W., W. Lang, M. S. Ascher, E. Vittinghoff, and W. Winkelstein (1993), The characterization of non-progressors: long-term HIV-1 infection with stable CD4+ T-cell levels, *AIDS*, *7*(9), 1159–1166.

Shi, L., et al. (2006), The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements, *Nat. Biotechnol.*, *24*, 1151–1161.

Shi, L., et al. (2010), The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models, *Nat Biotechnol*, *28*, 827–838.

Shipp, M. A., et al. (2002), Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, *Nat. Med.*, *8*, 68–74.

Shippy, R., et al. (2006), Using RNA sample titrations to assess microarray platform performance and normalization techniques, *Nat. Biotechnol.*, *24*(9), 1123–1131.

Showe, M. K., et al. (2009), Gene expression profiles in peripheral blood mononuclear cells can distinguish patients with non-small cell lung cancer from patients with nonmalignant lung disease, *Cancer Res.*, *69*, 9202–9210.

Silva, F. P., S. M. Swagemakers, C. Erpelinck-Verschueren, B. J. Wouters, R. Delwel, H. Vrieling, P. van der Spek, P. J. Valk, and M. Giphart-Gassler (2009), Gene expression profiling of minimally differentiated acute myeloid leukemia: M0 is a distinct entity subdivided by RUNX1 mutation status, *Blood*, *114*(14), 3001–3007.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Sima, C., and E. R. Dougherty (2006), What should be expected from feature selection in small-sample settings, *Bioinformatics*, *22*(19), 2430–2436.

Simon, R. (2003), Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data, *Br. J. Cancer*, *89*, 1599–1604.

Simon, R. (2008), The use of genomics in clinical trial design, *Clin. Cancer Res.*, *14*(19), 5984–5993.

Simon, R., M. D. Radmacher, K. Dobbin, and L. M. McShane (2003), Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification, *J. Natl. Cancer Inst.*, *95*, 14–18.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer (2005), ROCR: visualizing classifier performance in R, *Bioinformatics*, *21*(20), 3940–3941.

Sinnaeve, P. R., et al. (2009), Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease, *PLoS One*, *4*, e7037–e7037.

Skog, J., et al. (2008), Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers, *Nat. Cell Biol.*, *10*(12), 1470–1476.

Slonim, D. K. (2002), From patterns to pathways: gene expression data analysis comes of age, *Nat. Genet.*, *32 Suppl*, 502–508.

Smyth, G. K. (2004), Linear models and empirical bayes methods for assessing differential expression in microarray experiments, *Statistical Applications in Genetics and Molecular Biology*, *3*(1), Article 3.

Smyth, G. K., and T. Speed (2003), Normalization of cDNA microarray data, *Methods*, *31*, 265–273.

Sorlie, T., C. M. Perou, and Tibshirani, R. et al. (2001), Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, *PNAS USA*, *98*(19), 10,869–10,874.

Sotiriou, C., and M. J. Piccart (2007), Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?, *Nat. Rev. Cancer*, *7*, 545–553.

Speed, T. (2003), *Statistical Analysis of Gene Expression Microarray Data*, Chapman & Hall/CRC, FL,USA.

Staratschek-Jox, A., S. Classen, A. Gaarz, S. Debey-Pascher, and J. L. Schultze (2009), Blood-based transcriptomics: leukemias and beyond, *Expert Rev. Mol. Diagn.*, *9*, 271–280.

Statnikov, A., C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy (2005), A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, *Bioinformatics*, *21*(5), 631–643.

Stirewalt, D. L., et al. (2008), Identification of genes with abnormal expression changes in acute myeloid leukemia, *Genes Chromosomes Cancer*, *47*, 8–20.

Struyf, J., S. Dobrin, and D. Page (2008), Combining gene expression, demographic and clinical data in modeling disease: a case study of bipolar disorder and schizophrenia, *BMC Genomics*, *9*, 531.

Su, Z., B. Ning, H. Fang, H. Hong, R. Perkins, W. Tong, and L. Shi (2011), Next-generation sequencing and its applications in molecular diagnostics, *Expert Rev. Mol. Diagn.*, *11*(3), 333–343.

Swets, J. A. (1988), Measuring the accuracy of diagnostic systems, *Science*, *240*, 1285–1293.

Tan, K. S., A. Armugam, S. Sepramaniam, K. Y. Lim, K. D. Setyowati, C. W. Wang, and K. Jeyaseelan (2009), Expression profile of MicroRNAs in young stroke patients, *PLoS ONE*, *4*(11), e7689.

Tan, P. K., T. J. Downey, E. L. Spitznagel, P. Xu, D. Fu, D. S. Dimitrov, R. A. Lempicki, B. M. Raaka, and M. C. Cam (2003), Evaluation of gene expression measurements from commercial microarray platforms, *Nucleic Acids Res.*, *31*, 5676–5684.

Tarca, A. L., V. J. Carey, X. W. Chen, R. Romero, and S. Draghici (2007), Machine learning and its applications to biology, *PLoS Comput. Biol.*, *3*, e116.

Taylor, I. W., et al. (2009), Dynamic modularity in protein interaction networks predicts breast cancer outcome, *Nat. Biotechnol.*, *27*(2), 199–204.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2002), Diagnosis of multiple cancer types by shrunken centroids of gene expression, *PNAS*, *99*(10), 6567–6572.

Tillinghast, G. W. (2010), Microarrays in the clinic, *Nat. Biotechnol.*, *28*(8), 810–812.

Tseng, G. C., D. Ghosh, and E. Feingold (2012), Comprehensive literature review and statistical considerations for microarray meta-analysis, *Nucleic Acids Res.*, *40*(9), 3785–3799.

Tsujinishi, D., and S. Abe (2003), Fuzzy least squares support vector machines for multiclass problems, *Neural Netw*, *16*(5-6), 785–792.

Tusher, V. G., R. Tibshirani, and G. Chu (2001), Significance analysis of microarrays applied to the ionizing radiation response, *Proc. Natl. Acad. Sci. U.S.A.*, *98*, 5116–5121.

Twine, N. C., et al. (2003), Disease-associated expression profiles in peripheral blood mononuclear cells from patients with advanced renal cell carcinoma, *Cancer Res.*, *63*, 6069–6075.

Urbanke, J. (2012), http://cran.r-project.org/web/packages/multicore/multicore.pdf.

Vahey, M. T., et al. (2010), Expression of genes associated with immunoproteasome processing of major histocompatibility complex peptides is indicative of protection with adjuvanted RTS,S malaria vaccine, *J. Infect. Dis.*, *201*(4), 580–589.

Valk, P. J. M., et al. (2004), Prognostically useful gene-expression profiles in acute myeloid leukemia, *N Engl J Med*, *350*, 1617–1628.

van de Vijver, M. J., et al. (2002), A gene-expression signature as a predictor of survival in breast cancer, *N. Engl. J. Med.*, *347*, 1999–2009.

Van den Bulcke, T., K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor, and K. Marchal (2006), SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms, *BMC Bioinformatics*, *7*, 43.

van Iterson, M., P. A. 't Hoen, P. Pedotti, G. J. Hooiveld, J. T. den Dunnen, G. J. van Ommen, J. M. Boer, and R. X. Menezes (2009), Relative power and sample size analysis on gene expression profiling data, *BMC Genomics*, *10*, 439.

Van Looy, S., T. Verplancke, D. Benoit, E. Hoste, G. Van Maele, F. De Turck, and J. Decruyenaere (2007), A novel approach for prediction of tacrolimus blood concentration in liver transplantation patients in the intensive care unit through support vector regression, *Crit Care*, *11*, R83.

van 't Veer, L. J., et al. (2002), Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, *415*, 530–536.

Vapnik, V. N. (1998), *Statistical Learning Theory*, Wiley, NY,USA.

Varma, S., and R. Simon (2006), Bias in error estimation when using cross-validation for model selection, *BMC Bioinformatics*, *7*, 91.

Walter, S. D. (2005), The partial area under the summary ROC curve, *Stat Med*, *24*(13), 2025–2040.

Wang, S. J., and J. J. Chen (2004), Sample size for identifying differentially expressed genes in microarray experiments, *J. Comput. Biol.*, *11*(4), 714–726.

Wang, W. X., B. R. Wilfred, D. A. Baldwin, R. B. Isett, N. Ren, A. Stromberg, and P. T. Nelson (2008), Focus on RNA isolation: obtaining RNA for microRNA (miRNA) expression profiling analyses of neural tissue, *Biochim. Biophys. Acta*, *1779*(11), 749–757.

Watford, W. T., et al. (2008), Tpl2 kinase regulates T cell interferon-gamma production and host resistance to Toxoplasma gondii, *J Exp Med*, *205*, 2803–2812.

Watson, J. D., and F. H. Crick (1953), Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid, *Nature*, *171*, 737–738.

Westfall, P., and S. Young (1993), *Resampling-Based Multiple Testing, Examples and Methods for p-Value Adjustment.*, Wiley, New York USA.

World Health Organization (2012), http://www.who.int/.

Yang, M. C., J. J. Yang, R. A. McIndoe, and J. X. She (2003), Microarray experimental design: power and sample size considerations, *Physiol. Genomics*, *16*(1), 24–28.

Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed (2002a), Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Res.*, *30*, e15.

Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed (2002b), Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Res.*, *30*(4), e15.

Yeang, C. H., et al. (2001), Molecular classification of multiple tumor types, *Bioinformatics*, *17 Suppl 1*, S316–S322.

Yeoh, E.-J., et al. (2002), Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, *Cancer Cell*, *1*, 133–143.

Youden, W. J. (1950), Index for rating diagnostic tests, *Cancer*, *3*, 32–35.

Zander, T., et al. (2011), Blood-based gene expression signatures in non-small cell lung cancer, *Clin Cancer Res*, *17*, 3360–3367.

Zhan, F., et al. (2006), The molecular classification of multiple myeloma, *Blood*, *108*(6), 2020–2028.

Zhang, X., et al. (2006), Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data, *BMC Bioinformatics*, *7*, 197.