Using whole-genome wide gene expression profiling for the establishment of RNA fingerprints — application to scientific questions in molecular biology, immunology and diagnostics

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Daniela Eggle

aus

Neu-Ulm

Bonn, Februar 2008

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

1. Gutachter: Prof. Dr. Joachim Schultze

2. Gutachter: Prof. Dr. Jürgen Bajorath

Tag der Promotion: 05.05.2008

Diese Dissertation ist auf dem Hochschulschriftenserver der ULB Bonn http://hss.ulb.uni-bonn.de/diss_online elektronisch publiziert.

Erscheinungsjahr: 2008

Zusammenfassung

In der folgenden Arbeit wird der Begriff eines „RNA fingerprints“ eingeführt und an verschiedenen Fragestellungen der Immunologie und medizinischen Diagnostik angewendet. Unter einem „RNA fingerprint“ versteht man transkriptionelle Veränderungen, die durch ein molekulares Signal hervorgerufen werden. Dieses molekulare Signal kann ein aktivierter Signalweg, hervorgerufen durch Behandlung einer Zelle sein oder eine Krankheit, die bestimmte, für diese Krankheit spezifische, transkriptionelle Veränderungen hervorruft. In dieser Arbeit werden vier verschiedene Konzepte eines „RNA fingerprints“ vorgestellt. Das erste Konzept befasst sich mit *in vitro* definierten „RNA fingerprints“. „RNA fingerprints“ von T-Zell-inhibitorischen Molekülen, unter anderem TGFβ und PD-1, wurden *in vitro* erstellt und dann unter Verwendung bioinformatischer Methoden in Patienten mit Hodgkin Lymphom nachgewiesen. Somit konnten wir zeigen, dass T-Zellen im Tumormilieu des Hodgkin Lymphoms unter dem Einfluss von TGFβ und PD-1 stehen. Im zweiten Konzept wird die Erstellung eines krankheitsspezifischen „RNA fingerprints“ vorgestellt. Mit Hilfe von Transkriptionsprofilen von Lungenkrebs- und Kontrollpatienten wird hier ein Bronchialkarzinom-spezifischer „RNA fingerprint“ erstellt, welcher das Auftreten einer Lungenkrebserkrankung, welche noch nicht klinisch manifest ist, voraussagen kann. Ein weiteres Konzept befasst sich mit der Verwendung von vordefinierten „RNA fingerprints“. Diese können aus biologischen Datenbanken extrahiert werden und umfassen Gene, welche spezifischen Signal- oder Stoffwechselwegen zugehörig sind. Ich habe eine neue, sehr einfache „Gene class testing“-Methode entwickelt, welche vordefinierte „RNA fingerprints“ aus der Gene Ontology testet. Das letzte Konzept befasst sich mit der Idee, das Microarray Experiment als solches als “RNA fingerprint” zu betrachten. Ich stellte die Hypothese auf, dass alle transkriptionellen Veränderungen eines Experiments als „RNA fingerprint“ dieses Experiments betrachtet werden können. Hierfür wurde die „Gene-class testing“-Methode um einen Netzwerk-Algorithmus erweitert, um Spielmacher-Moleküle in diesem Experiment zu identifizieren. Des Weiteren wird in der Arbeit ein Softwarepaket vorgestellt, welches es Wissenschaftlern ermöglicht, die Konzepte der „RNA fingerprints“ anzuwenden. Aufgrund einer Grafik-basierten Benutzeroberfläche können Microarray-Experimente analysiert werden, ohne dass Programmierkenntnisse erforderlich sind. Essentiell für die Durchführung und Anwendung aller vorgestellten „RNA fingerprint“-Konzepte ist die Verlässlichkeit der zugrundeliegenden Technologie, in diesem Fall der Microarray-Technologie. Am Beispiel der Illumina BeadChip Technologie wird zum Abschluss kritisch beurteilt, inwieweit die Veränderung eines Chips (Technologie-Veränderungen, Inhalt des Arrays) Einfluss auf die erzielten Resultate hat.

Preface

The following thesis covers the main part of research generated during my doctorate studies in the Molecular Tumor Biology and Tumor Immunology group of Prof. Dr. Joachim Schultze at the University Clinics of Cologne from August 2005 to January 2008. I am very grateful to Prof. Dr. Joachim Schultze for giving me the opportunity to work in this stimulating and exciting field at the border of molecular biology, immunology, medicine and informatics. It was especially his enthusiasm and his gratification to work that contributed to the success of this thesis and my dedication in this field of research.

I would also like to thank very much the people in my group: Dr. Svenja Debey-Pascher, with whom I worked in close collaboration on almost every project. Sabine Classen, who I could count on in any possible situation, either with problems at work or private. Julia Driesen, Dr. Marc Beyer and Dr. Alexey Popov for fruitful discussions and their support concerning biological and medical questions. I would also like to thank the three technicians Julia, Mirela and Ingrid. I did not work with them directly, but without their dedicated work I would not have had any data to work on. I have to thank the other House 16UG people, especially Luise Gryschok, Michaela Patz and Tanja Liebig who made the UG a fun place to work. Finally I would like to thank our clinicians Dr. Jens Chemnitz and Dr. Thomas Zander for a good collaboration throughout my thesis. Especially the discussions with Dr. Thomas Zander concerning statistics and bioinformatics very much contributed to the success of my work. And a more private note, I would like to thank Peter for his encouragement and especially his understanding in busy times. Also I would like to thank my parents for their continuous and dedicated support during my years of study.

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

Chapter 4 introduces the term of RNA fingerprints in its original form applied to the immunological problem of immune inhibition within the tumor environment. The study was performed in close collaboration with Dr. Jens Chemnitz, who introduced the idea of interrogating the contribution of different inhibitory molecules to the tumor environment of Hodgkin's disease. All considerations about stating this problem in the context of RNA fingerprints and all calculations were performed by me. In July 2007, the results of this study have been published in *Blood* (Chemnitz, Eggle et al. 2007).

Chapter 5 covers the concept of RNA fingerprints in a diagnostic setting. The research on this topic was performed in close collaboration with Dr. Thomas Zander and Dr. Svenja Debey-Pascher. The results of this study are currently prepared as a manuscript for submission.

Chapter 6 introduces a new gene-class testing approach (GOAna) to test pre-defined RNA fingerprints for their contribution to changes between subgroups in a microarray experiment. The algorithm is based on a preliminary idea of Dr. Benedikt Brors (DKFZ Heidelberg) and Dr. Thomas Zander and has been implemented and enhanced by me. The application of GOAna to the immunological problem of T cell homeostasis was performed in close collaboration with Sabine Classen. The bioinformatics part of this study, including the application of the implemented approach was carried out by me. In June 2007 the results of this study have been published in the *Journal of Immunology* (Classen, Zander et al. 2007)

Chapter 7 expands the GOAna algorithm and aims for the identification of key players within a microarray experiment by using the measured transcriptional changes as a RNA fingerprint. The algorithm was designed and implemented by me and was further applied to an unsolved biological question in T cell biology, the unraveling of detailed signaling mechanisms following inhibition of T cells. All biological experiments for substantiating the findings in this study were carried out by Julia Driesen. The results have been presented as a poster at the ISMB/ECCB 2007 in Vienna, Austria and are prepared for publication.

Chapter 8 introduces a software application (IlluminaGUI) which is intended to help researchers, especially non-bioinformaticians to analyze microarray data derived from the Illumina BeadChip platform. Idea, design and implementation of this software package were performed by me. Dr. Svenja Debey-Pascher performed the beta-tests on the software. In June 2007 the results have been published in *Bioinformatics* (Schultze and Eggle 2007)

Chapter 9 deals with critical considerations about the technology underlying the determination of RNA fingerprints. All calculations and investigations in this project were performed by me. Dr. Svenja Debey-Pascher supported this research by calculations of the whole blood data set introduced in this chapter. The results of this study have been submitted for publication.

Table of contents

List of Figures

List of Tables

Part I: General Introduction

1 Introduction

This thesis focuses on the development of different concepts of RNA fingerprints on the basis of transcriptional profiling using microarrays. DNA microarrays are the major technology used for establishing genome-wide transcriptional profiles of cells, tissues or even whole organs (Schena, Shalon et al. 1995). With the introduction of this technology, researchers have started to describe changes in gene expression between different samples (Schena, Shalon et al. 1995; DeRisi, Penland et al. 1996; Lockhart, Dong et al. 1996; Spellman, Sherlock et al. 1998) including the determination of differentially expressed genes and the grouping of genes based on their expression pattern across samples using unsupervised classification methods. Additionally supervised classification methods have been used to systematically classify diseases based on transcriptional changes (Golub, Slonim et al. 1999; Alizadeh, Eisen et al. 2000; Shipp, Ross et al. 2002; Valk, Verhaak et al. 2004). Several different algorithms have been introduced, all aiming on the sub-classification, prediction and diagnosis of different diseases (Vapnik 1998; Tibshirani, Hastie et al. 2002). The term RNA fingerprint has been introduced in our lab and can be assigned to generally all predictive gene signatures which are generated from transcriptional profiles that are based on biological differences between sample groups. Starting out with a concept introduced by the Nevins' lab in 2003 in which the group created predictive gene signatures for different oncogenes *in vitro* and demonstrated the existence of these molecules *in vivo,* I hypothesized that - in principle - this concept should be applicable to any other molecular factor that leads to transcriptional changes upon stimulation and signaling. That means, observed transcriptional changes are biological responses of any given cell in reply to a molecular signal and can therefore be termed a RNA fingerprint of the respective signal. Molecular signals include activated oncogenic pathways by introducing the oncogene as a transgene, receptor ligand interactions, treatment of cells with inhibitory factors and responses of cells to different diseases. Here I give a short overview of the different concepts of RNA fingerprints that are described within this thesis.

The thesis is structured as follows:

In *Chapter 2* I will present all background knowledge that is needed to understand and follow this whole thesis. That means that all biological, technical and bioinformatics terms which are mentioned throughout the thesis are clarified in this section. The fundamentals of molecular biology and immunology will be explained as well as genome-wide transcriptional profiling using microarray

technology with a special focus on the recently introduced Illumina BeadChip technology. Finally, several data analysis methods that are applied in Part II and Part III of the thesis will be introduced. Most of the presented material is textbook knowledge and can also be found in (Alberts, Bray et al. 2002; Speed 2003; Gentleman, Carey et al. 2005; Janeway, Travers et al. 2005; Crawley 2007).

Chapter 3 briefly introduces the most important Material and Methods used for this thesis. Since this thesis focuses on the bioinformatics part of the studies introduced here, a detailed description of the Material and Methods used for the experimental setups can be found in (Driesen 2005; Chemnitz, Eggle et al. 2007; Classen 2008). An overview of the experimental methods used can be found in Appendix C.

In *Chapter 4* we generated gene signatures for different T cell inhibitory molecules, including TGFβ and PD-1, *in vitro* and introduced these signatures as RNA fingerprints of the molecules. These fingerprints are then applied to gene expression profiles of human cancers to directly determine the *in vivo* impact of the interrogated molecules on tumor infiltrating T cells. By applying supervised and unsupervised classification methods based on the RNA fingerprints of both, TGFβ and PD-1 it was then shown that T cells derived from patients with Hodgkin's lymphoma are indeed under the influence of both, TGFβ and PD-1.

Chapter 5 extends the concept to a disease specific RNA fingerprint in a diagnostic setting. Transcriptional changes which are an image for the disease should be able to specifically distinguish this disease not only from healthy controls, but also from any other disease and can therefore be termed RNA fingerprint for this disease. In this chapter a lung cancer specific RNA fingerprint was developed to predict the occurrence of lung cancer prior to clinical manifestation.

Chapter 6 introduces a further concept which deals with the use of pre-defined RNA fingerprints. These can be extracted from biological databases that include information about genes belonging to special pathways or groups of genes with similar functions. I have developed a new and very simple gene-class testing method, GOAna, which is based on RNA fingerprints provided by the Gene Ontology (GO) Consortium. Using GOAna, it is possible to perform an unbiased analysis based on all branches of GO.

In *Chapter 7* the fourth and last concept introduces the idea of using the microarray experiment itself as a RNA fingerprint. I hypothesized that all transcriptional changes which are revealed by a

microarray experiment can serve as a RNA fingerprint and can decipher underlying signaling mechanisms. The algorithm presented in Chapter 6 was extended by a network-construction algorithm to determine key player genes which link the identified significant gene spaces. Using this approach a key player within the PGE_2 signaling pathway in $CD4^+$ T cells was identified and experimentally validated.

Chapter 8 introduces a software package, IlluminaGUI, which allows the researcher to establish and apply RNA fingerprints to gene expression data derived from Illumina's Sentrix BeadChip technology. IlluminaGUI is implemented as a graphical user interface and is intended to enable the interested life scientist who is not familiar with a command line based environment like the R language to analyze microarray experiments.

In *Chapter 9* critical issues concerning the used technology are raised. All described approaches for the creation of RNA fingerprints are heavily dependent on the reliability of the microarray format used for the study. Here the continuity of RNA fingerprints is discussed when a new version of a microarray with updated probe content becomes available.

The last part of the thesis gives a summary of all previous discussions and aligns these discussions into a broader context. Here future research directions are pointed out and possible difficulties concerning the concept of RNA fingerprints are raised.

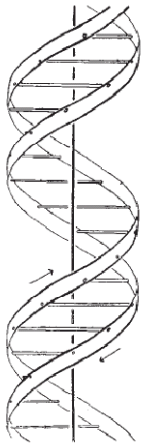
2 Background information

In this chapter I would like to introduce the reader to all background knowledge that is needed to understand and follow this whole thesis. Additionally all biological, technical and bioinformatics terms which are mentioned throughout the thesis are clarified in this section. First I sketch the fundamentals of molecular biology and immunology, starting with the definition of basic terms like DNA, RNA and proteins and concluding with an overview over the human immune system and basic immunological methods for immune cell extraction. Then I will briefly introduce the fundamentals of cancer and present three examples of cancer types in more detail. Furthermore I will introduce the reader to genome-wide transcriptional profiling using microarray technology with a special focus on the recently introduced Illumina BeadChip technology. Finally, several data analysis methods that are applied in Part III of the thesis will be introduced.

2.1 A primer on molecular biology

2.1.1 DNA

The history of DNA goes back to 1868 when a young Swiss scientist called Friedrich Miescher isolated a new substance from cell nuclei which he called nucleic acid (Dahm 2005). That this substance holds the genetic information of the cell was discovered by Oswald T. Avery in 1943 (Avery, MacLeod et al. 1979). In 1953, Watson and Crick determined the spatial structure of DNA to be a double helix. Depicted on the left is the original figure derived from their article in Nature (Watson and Crick 1974). DNA is a very long, threadlike macromolecule arranged in two strings which are antidromic and is made up of a large number of deoxyribonucleotides, each composed of an organic base (adenine, guanine, cytosine or thymine), a sugar (Pentose) and a phosphate group. The bases of the DNA molecules carry genetic information whereas their sugar and phosphate groups perform a structural role.



2.1.2 RNA

Although DNA holds the genetic information, it is not the direct template for protein synthesis. The direct template for protein synthesis is the RNA molecule, a long, unbranched macromolecule which,

like DNA, consists of nucleotides. In contrast to DNA, the sugar unit in RNA is ribose and thymine is replaced by the derivate uracil. Also, RNA molecules are usually single-stranded, except in some viruses. There are two major types of RNA, coding and non-coding RNA. **Table 2.1** lists the different RNA molecules together with their function.

RNA name	RNA type	Function
Messenger RNA (mRNA)	coding	Template for protein synthesis
Transfer RNA (tRNA)	non-coding	Translation
Ribosomal RNA (rRNA)	non-coding	Translation
Antisense RNA (aRNA)	non-coding	Gene regulation
MicroRNA (miRNA)	non-coding	Gene regulation
Small interfering RNA (siRNA)	non-coding	Gene regulation

Table 2.1 – Types of RNA molecules
There are two major types of RNA molecules, coding and non-coding RNA. RNA name (abbreviation), type and function are depicted.

Messenger RNA (mRNA) is a coding RNA which serves as the template for protein synthesis. Non-coding RNA genes are genes that encode RNA which is not translated into a protein. The most prominent representatives of non-coding RNAs are transfer RNAs (tRNA) and ribosomal RNAs (rRNA). tRNA carries amino acids in an activated form to the ribosome for peptide-bond formation, in a sequence determined by the mRNA template. rRNA, the major component of ribosomes, plays both a catalytic and structural role in protein synthesis. Other non-coding RNAs include antisense RNA (aRNA), microRNA (miRNA) and small interfering RNA (siRNA) which all function as gene regulation molecules.

2.1.3 Proteins

Proteins play crucial roles in virtually all biological processes. For instance, they facilitate biochemical reactions, transfer signals, function as antibodies in the immune system, and actively transport other molecules. Structurally, proteins are linear polymers built from 20 different amino acids, all of which have a common base structure to which a specific side chain is attached; they are linked together by peptide bonds. The side chains are critical for the function of a protein because they can have many different chemical properties, for example, they can differ in size, shape, charge or chemical

reactivity; the arrangement of different amino acids therefore lends a protein its function through a specific combination of these chemical properties. Each protein has its own unique amino acid sequence that is specified by the nucleotide sequence of the gene encoding this protein.

2.1.4 From DNA to RNA to protein

The flow of genetic information in normal cells is from DNA to RNA to protein. The synthesis of RNA from a DNA template is called transcription, whereas the synthesis of protein from an RNA template is termed translation. The relation between the sequence of bases in DNA (or its mRNA transcript) and the sequence of amino acids in a protein is called the genetic code. The code is nearly the same in all organisms and defines a mapping between tri-nucleotide sequences called codons and amino acids; every triplet of nucleotides in a nucleic acid sequence specifies a single amino acid.

2.2 A primer on immunology

2.2.1 Innate vs. adaptive immune system

The immune system is composed of two major subdivisions, the innate or nonspecific immune system and the adaptive or specific immune system. Although both systems function to protect against invading organisms, they differ in a number of ways. In the early phases of the host response to infection the cells of the innate immune system recognize and respond to pathogens in a generic way. The innate immune system is therefore the first line of defense against invading organisms, since most cells are constitutively present and ready to be mobilized upon infection. The adaptive immune system, on the other hand, requires some time to react to an invading organism since it is composed of highly specialized cells and processes that eliminate pathogenic challenges. It is antigen specific and reacts only with the organism that induced the response. In contrast, the innate system is not antigen specific and reacts equally well to a variety of organisms, and does not discriminate between pathogens. Finally, the adaptive immune system demonstrates immunological memory. It “remembers” that it has encountered an invading organism and reacts more rapidly on subsequent exposure to the same organism. In contrast, the innate immune system does not demonstrate immunological memory and does not increase with repeated exposure.

2.2.2 The cells of the immune system

Both innate immunity and adaptive immunity responses depend upon the activities of white blood cells or leukocytes. Leukocytes are found throughout the body, including the blood and lymphatic system. Several different types of leukocytes exist, but they all derive from a pluripotent cell in the bone marrow, the hematopoietic stem cell (**Figure 2.1**).

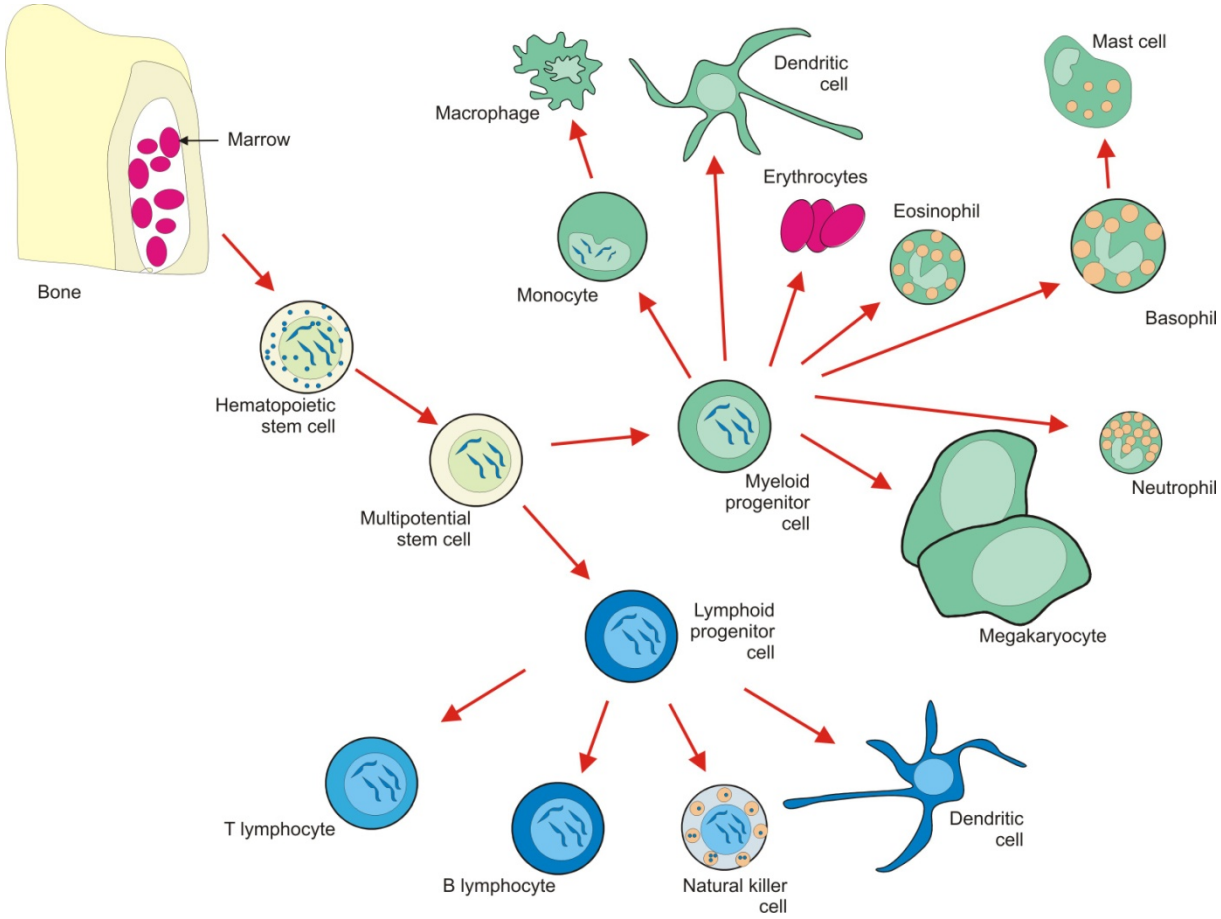


Figure 2.1 – All cells of the immune system arise from hematopoietic stem cells in the bone marrow
The pluripotent hematopoietic stem cells divide to produce several types of progenitor cells, the lymphoid stem cell (lymphoid progenitor cell) and the myeloid stem cell (myeloid progenitor cell). The lymphoid progenitor cell gives rise to lymphocytes, including T lymphocytes (T cells), B lymphocytes (B cells), natural killer cells (NK cells) and dendritic cells (DCs). The myeloid progenitor cell gives rise to, for example, monocytes, macrophages, eosinophils or basophils.

The pluripotent hematopoietic stem cells divide to produce several types of progenitor cells, including lymphoid progenitor cells and myeloid progenitor cells. The myeloid progenitor cells develop into the cells that respond early and nonspecifically to infection, i.e. cells which are part of the innate immune system. Macrophages and neutrophils are primarily phagocytic cells that engulf bacteria upon contact and send out warning signals. Eosinophils are involved in attacking parasites,

while basophils release the contents of their granules containing allergy-related molecules. The lymphoid progenitor cells develop into the small white blood cells called lymphocytes. Lymphocytes are the major component of the adaptive immune system and include two main classes, the B lymphocytes (B cells) and the T lymphocytes (T cells). Upon 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. A subset of T cells, the $CD4^+$ T cells, will be used in the experiments throughout the thesis. This subset is introduced in more detail below.

2.2.2.1 T cells

T cells are a subset of lymphocytes which work at the core of adaptive immunity. The abbreviation T, in T cell, stands for thymus, the principal organ in a T cell's development. T cells are usually divided into two major subsets that are functionally and phenotypically different, the cytotoxic T cells and the helper T cells. The cytotoxic T cells, or killer T cells, eliminate cells which are infected with parasites as well as cells that have been transformed by cancer but have not yet adapted to evade the immune detection system. They are also responsible for the rejection of tissue and organ grafts. Cytotoxic T cells are activated when their T cell receptor (TCR) recognizes a specific antigen presented by another cell. This recognition is aided by a co-receptor on the T cell, called CD8, hence the name $CD8^+$ T cells.

The helper T cells, also called $CD4^+$ T cells, are coordinators of immune regulation. $CD4^+$ T cells are also activated by recognizing a specific antigen on an antigen-presenting cell. But these cells have no cytotoxic activity and do not kill infected cells directly. Instead activation of a $CD4^+$ T cell causes it to release cytokines that influence the activity of many immune cells and therefore, for example, enhances the activity of cytotoxic T cells. In addition, activation of $CD4^+$ T cells leads to an up-regulation of different molecules expressed on the T cell's surface, including CD40 ligand, which provide extra stimulatory signals required to activate antibody-producing B cells.

2.2.3 Isolation of lymphocytes

Human lymphocytes can be isolated from peripheral blood by density gradient centrifugation using the polymer Ficoll. In short, peripheral blood is layered over Ficoll and is centrifuged. Red blood cells and polymorphonuclear leukocytes or granulocytes are more dense than mononuclear cells and centrifuge through the Ficoll. This yields a population of mononuclear cells (peripheral blood

mononuclear cells (PBMC)) at the interface that consists mainly of lymphocytes and monocytes. In experimental animals, and occasionally in humans, lymphocytes can also be isolated from lymphoid organs, such as spleen, thymus, bone marrow or lymph nodes.

For the isolation of a particular cell population from a sample or culture many different methods exist, including magnetic cell separation (MACS). Here, the cells are incubated with magnetic beads coated with antibodies against a particular surface antigen. Cells expressing this antigen attach to the magnetic beads, while cells not expressing the antigen flow through. With this method, the cells can be separated positively or negatively with respect to the particular antigen(s).

For the isolation of very rare or highly-purified cell populations a fluorescence-activated cell sorter (FACS) can be used. Individual cells within a mixed population are first tagged by a fluorescently labeled antibody. The cells are then forced through a nozzle in a single-cell stream that passes through a laser beam. Photomultiplier tubes (PMTs) detect the scattering of light, a sign of cell size and granularity, and emissions from the different fluorescent dyes. In this way, specific subpopulations of cells, distinguished by the binding of the labeled antibody, can be purified from a mixed population of cells.

2.3 A primer on cancer

Cancer is a generic term for a group of more than 100 diseases that can affect any part of the body. It is defined by a rapid creation of abnormal cells which grow beyond their usual boundaries, and which can invade adjoining parts of the body and spread to other organs, a process known as metastasis. Cancer is a leading cause of death worldwide (Parkin, Bray et al. 2005). From a total of 58 million deaths worldwide in 2005, cancer accounts for 7.6 million (or 13%) of all deaths (World Health Organization 2008).

Cancers are classified by the type of cell that resembles the tumor and, therefore, the tissue presumed to be the origin of the tumor. Two examples of general categories include:

- Carcinoma: Malignant tumors derived from epithelial cells. This group represents the most common cancers, including the common forms of breast, prostate, lung and colon cancer.
- Lymphoma and leukemia: Malignancies derived from hematopoietic (blood-forming) cells. Examples include acute myeloid leukemia, Hodgkin's lymphoma and follicular lymphoma.

In the course of this thesis three different types of cancer are addressed within the introduced studies; Hodgkin's lymphoma, follicular lymphoma and lung cancer. Below is a short introduction to these three diseases.

2.3.1 Lung cancer

Lung cancer is a disease of uncontrolled cell growth in tissues of the lung. This growth may lead to metastasis, invasion of adjacent tissue and infiltration beyond the lungs. The vast majority of primary lung cancers are carcinomas of the lung, derived from epithelial cells. Lung cancer is the most common cause of cancer-related death in men and the second most common in women (Parkin, Bray et al. 2005) and is responsible for 1.3 million deaths worldwide annually. Symptoms of lung cancer include shortness of breath, hoarseness, chronic fatigue, loss of appetite and unexplained weight loss (World Health Organization 2008). Lung cancer is classified as two major types: small cell lung cancer (SCLC) and non–small cell lung cancer (NSCLC). SCLC accounts for 15% of all lung cancers and is an aggressive and fast-growing cancer that forms in tissues of the lung and can spread to other parts of the body. The cancer cells look small and oval-shaped when looked at under a microscope. NSCLC is the most common kind of lung cancer and includes a group of lung cancers. The three main types of non-small cell lung cancer are squamous cell carcinoma, large cell carcinoma, and adenocarcinoma (National Institute of Health 2008). Lung cancers are described in different stages, starting from an occult stage in which lung cancer cells are found in sputum or in a sample of water during bronchoscopy, but without a visible tumor in the lung to stage IV where malignant growths of cells may be found in more than one lobe of the same lung or in the other lung.

2.3.2 Hodgkin lymphoma and follicular lymphoma

Hodgkin lymphoma or Hodgkin's disease is a type of lymphoma which is characterized by the presence of Reed-Sternberg cells. The two major types of Hodgkin lymphoma are classical Hodgkin lymphoma and nodular lymphocyte-predominant Hodgkin lymphoma. Symptoms include the painless enlargement of lymph nodes, spleen, or other immune tissue. Other symptoms include fever, weight loss, fatigue, or night sweats. Treatment of Hodgkin lymphoma is performed using chemotherapy, radiation or stem cell transplantation (National Institute of Health 2008). Follicular lymphoma is a common type of Non-Hodgkin Lymphoma (NHL). It is a slow growing lymphoma that arises from B-cells and is therefore categorized as a B cell tumor. Symptoms include painless swelling in the neck, enlarged lymph nodes, fatigue and loss of appetite. As with Hodgkin lymphoma, follicular lymphoma is treated by radiation therapy, chemotherapy or monoclonal antibody therapy.

2.4 Microarray technology

The ability to assess genome-wide transcriptional profiles of cells, tissues or even whole organs is a cornerstone of the advances genomics has broad to the life and medical sciences (Pennacchio and Rubin 2001; Reinke and White 2002). DNA microarrays are the major technology used for this purpose (Schena, Shalon et al. 1995). Both in biology and medicine, important new findings have been revealed by this technology.

2.4.1 Basics

Microarray technology represents a powerful functional genomics technology which permits the expression profiling of thousands of transcripts in parallel. The technology is based on hybridization of complementary nucleotide strands (DNA or RNA). Microarray chips consist of thousands of DNA molecules that are immobilized and gridded onto a support such as glass, silicon or nylon membrane. Each spot on the chip is representative for a certain gene or transcript. Fluorescently or radioactively labeled nucleotides (targets) that are complementary to the isolated mRNA are prepared and hybridized to the immobilized molecules (probes). Targets that did not bind to probes during the hybridization process are washed away. The amount of hybridized target molecules is proportional to the amount of isolated mRNA. The relative abundance of hybridized molecules on a defined array spot can be determined by measuring the fluorescent or radioactive signal. This method provides the advantage that it can interrogate the level of transcription of several thousands of different genes from one sample in one experiment. Several competing technologies for microarray probe implementation have emerged, including the use of full-length cDNAs, or presynthesized or in situ synthesized oligonucleotides as probes. One of the “gold standard” technologies is the GeneChip distributed by Affymetrix. The GeneChips are a constructed using a combination of two techniques, photolithography and solid-phase DNA synthesis. Other distributors of DNA microarrays include GE Healthcare, Applied Biosystems, Beckman Coulter, Eppendorf Biochip Systems, Agilent and very recently Illumina.

2.4.2 The Illumina BeadChip system

In 2004, Illumina Inc. has developed a new microarray technology for quantitative gene-expression profiling. The technology completely differs from the Affymetrix system and is based on randomly assembled arrays of beads (**Figure 2.2**). Each glass slide is composed of six arrays each measuring

~50,000 transcripts. The probes used by Illumina are processed using standard oligonucleotide synthesis methods as used for spotted long-oligonucleotides arrays. However, the oligonucleotides are covalently attached to small microbeads (~700,000 copies of a particular oligonucleotide per bead) which are then put onto microarrays using a random self-assembly mechanism (Kuhn, Baker et al. 2004).

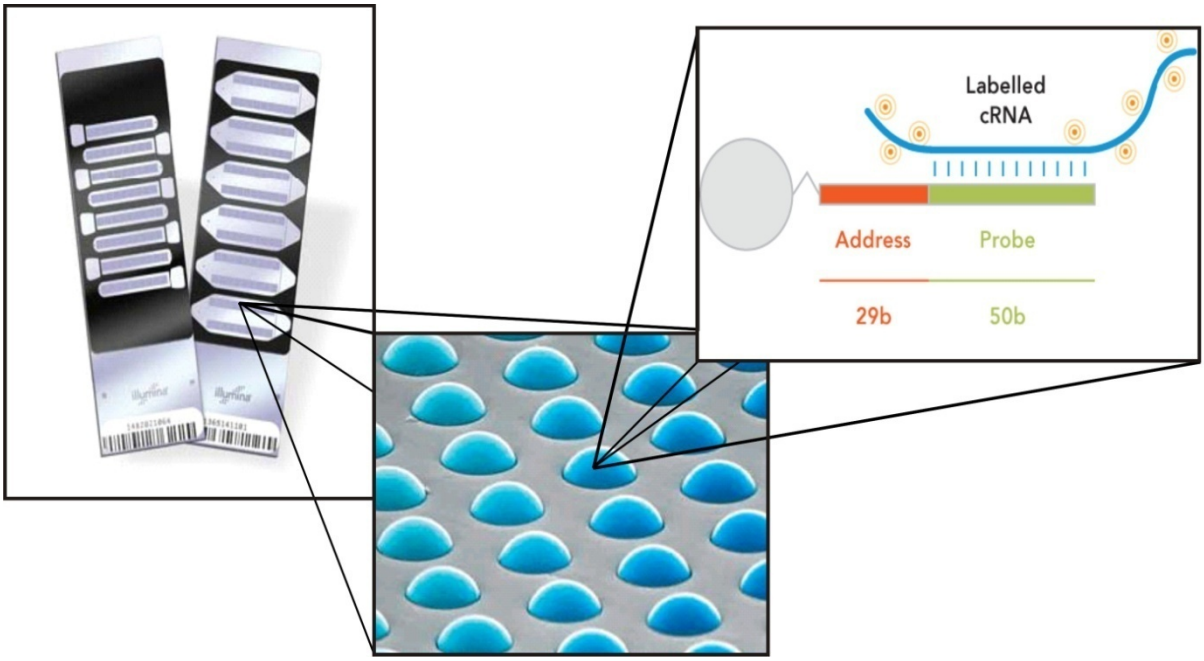


Figure 2.2 – Design of an Illumina BeadChip
Each glass slide is composed of six arrays each measuring ~5000 transcripts. The gene specific oligonucleotides are covalently attached to small microbeads together with an address sequence which is used to decode the position of the oligonucleotide on the array.

There are multiple copies of each sequence-specific bead on an array (on average 30 copies on any array), which contributes to measurement precision and reliability. Since the beads are randomly assembled on the array, each probe has associated with it an address sequence (29 base pairs, **Figure 2.2**). During the scanning process, the arrays undergo a decoding step (Gunderson, Kruglyak et al. 2004) in which this address sequence is used to determine the location of each probe on the array.

2.5 A primer on microarray data analysis methods

The analysis of DNA microarrays poses a large number of statistical problems, including the normalization of the data. A basic difference between microarray data analysis and much traditional

biomedical research is the dimensionality of the data. In all applications of microarray technology, the number of variables (transcripts) is much larger than the number of observations (chips): a typical study includes from 20000 to 50000 transcripts for only 10 to 200 chips. In contrast, a large clinical study might collect 100 data items per patient for thousands of patients. Based on this difference, adapted data analysis methods are required. Here, different methods which are used throughout this thesis are introduced.

2.5.1 Quality control

Quality control of microarray data is the first and probably one of the most important steps in a microarray analysis. There are different ways to examine the quality of microarray data, some of which are depicted here:

2.5.1.1 Basic diagnostic plots

Diagnostic plots include boxplots (**Figure 2.3A**), pairwise scatter plots (**Figure 2.3B**), and MA plots (introduced by Dudoit et al. (Dudoit, Yang et al. 2002), **Figure 2.3C**).

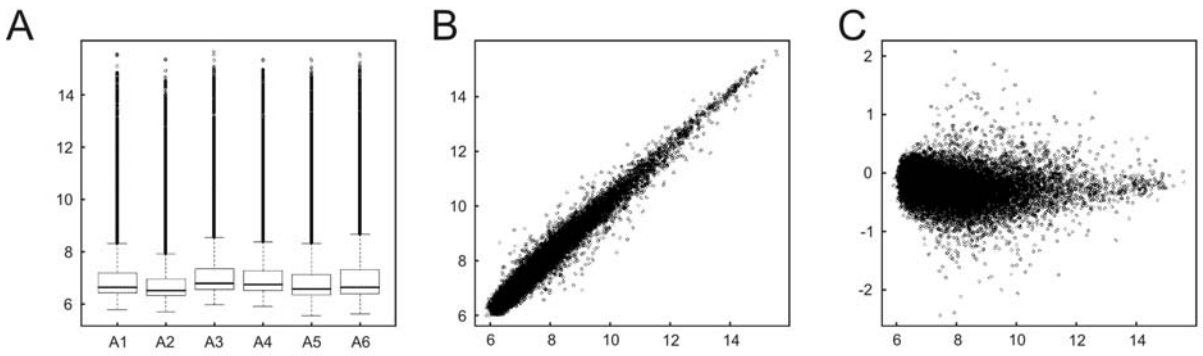


Figure 2.3 – Diagnostic plots for quality control
There are different diagnostic plot for quality control of microarray data. Depicted here are (A) a boxplot, (B) a pairwise scatter plot and (C) a MA plot.

A boxplot (also known as box-and-whisker diagram) is a way of graphically depicting groups of numerical data through their five-number summaries. These five numbers are (1) the smallest observation, (2) the lower quartile (Q1), (3) the median, (4) the upper quartile (Q3), and (5) the largest observation. The central box in the plot represents the inter-quartile range (IQR). The IQR is the range between the lower quartile value and the upper quartile value within which the middle

50% of the ranked data are found. Boxplots can be used to determine the distribution of intensity signals across an array, thereby verifying the comparability of all arrays within an experiment. A further common graphical display of microarray data is a scatterplot, in which the data is displayed as a collection of points, each having one coordinate on the horizontal axis and one on the vertical axis. A scatterplot therefore shows the linear relationship between variables. When, for example, performing a two-channel microarray experiment, the two channels (usually red and green) are plotted in a scatterplot to identify the relationship between dye-bias and signal intensity. When performing a one-channel experiment, biological replicates can be plotted to detect intensity-dependent differences. An MA plot is a rotation of the scatterplot by 45 degrees with a subsequent re-scaling of the data. For each gene, the fold-change M-value ($M = log_2(\mathrm{Array}_1/\mathrm{Array}_2)$) is plotted on the vertical axis and the intensity A-value ($A = \left(log_2(\mathrm{Array}_1) + log_2(\mathrm{Array}_2)\right)/2$) is plotted on the horizontal axis (**Figure 2.3C**). The MA plot therefore displays the relationship between differential expression and intensity and is used for comparing arrays from different groups.

2.5.1.2 Determination of absent and present status of probes

For each transcript on the microarray the scanning software determines both an expression signal and a detection p-value. The detection p-value is calculated by statistically comparing the expression signal to a negative control usually present on the microarray and therefore depicts a significance measure for the two signals being different. A probe is called present if the expression signal significantly differs from the negative signal, otherwise absent. The absent resp. present status of probes can be used to investigate sensitivity differences of arrays within an experiment. Here, the percentage of present probes on each array is calculated and compared to each other. Within an experiment, similar percentages for each array should be achieved.

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

2.5.2 Normalization

During a microarray experiment, different sources of systematic variation can affect the measured gene expression levels, including unequal quantities of starting material, differences in labeling or detection efficiencies between one experiment and the other. Usually, a normalization process is used to remove such variation from the data in order to detect biological differences between

samples. Many different normalization techniques have been implemented which are all based on different assumptions concerning the nature of the raw data. There has been and is still extensive research going on as to which normalization method performs best on which data. In my opinion, the most frequently used methods for one channel microarray data are the quantiles method (Bolstad, Irizarry et al. 2003), the vsn-method (Huber, von Heydebreck et al. 2002) and the qspline-method (Workman, Jensen et al. 2002). As already mentioned, depending on the data set, different normalization techniques give different results. Therefore, normalization techniques should be tested within an analysis and the best performing technique should then be used for further analysis.

2.5.3 Identification of differentially expressed genes

The first method used to evaluate whether a transcript shows different signal intensities between two groups, i.e. is differentially expressed, was to calculate a fold change (FC) between the two sample groups. To date however, the FC measure alone is considered as an inadequate test statistic because it does not incorporate variance and offers no associated level of 'confidence'. The biological question of differential expression was therefore restated as a problem in hypothesis testing: a test of the null hypothesis of no association between the expression levels and the responses. The different methods used for hypothesis testing within microarray data mainly differ in kind of test statistic used (e.g. parametric test statistic, non-parametric test statistic). In any testing situation, despite of the test statistic used, two types of errors can be committed (**Table 2.2**):

		Test result	
		p-value > α	p-value < α
Truth	no difference	√	type I error (false positive)
	difference	type II error (false negative)	√

Table 2.2 – Types of errors in hypothesis testing
Two types of errors can be committed in a testing situation, a type I error which is committed by declaring that a gene is differentially expressed when it is not, and a type II error, which is committed when the test fails to identify a truly differentially expressed gene.

A false positive, or Type I error, is committed by declaring that a gene is differentially expressed when it is not, and a false negative, or Type II error, is committed when the test fails to identify a truly differentially expressed gene. In case of a microarray experiment, the large number of

transcripts present on a single array represents a further problem, the problem of multiple testing: each transcript which is called significantly different between the two analyzed groups has a specified Type I error probability. The very high number of transcripts on an array multiplies this error probability and makes it likely that, just by chance, the differential expression of some transcripts represent false positives. There are different Type I error rates, including the family wise error rate (FWER) and the false discovery rate (FDR). The FWER is defined as the probability of at least one Type I error in the whole experiment. The FDR is the expected proportion of Type I errors among the rejected hypotheses. A number of methods have been established that address the question of multiple testing in microarray experiments and control a defined Type I error rate. The most widely used methods include significance analysis of microarrays (SAM) (Tusher, Tibshirani et al. 2001) which estimates the false discovery rate (FDR) and linear models for microarray analysis (LIMMA) (Smyth 2004) which uses adjusted p-values to control the FWER.

2.5.4 Classification

Classification is an important data analysis method for microarray experiments, for purposes of classifying biological samples and predicting clinical or other outcomes using gene expression data. One discriminates between unsupervised and supervised methods of classification.

Unsupervised classification, also known as cluster analysis or clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets or so-called clusters, such that the objects within each cluster are more closely related to one another than objects assigned to different clusters. There are different clustering approaches, including hierarchical clustering, k- means clustering, or clustering using Self Organizing Maps (SOMs). Hierarchical clustering, although or even because it is a very simple and intuitive concept, is one of the most widely used methods for unsupervised classification. It is described in more detail below.

In supervised classification, also known as class prediction, the class of each sample in the data set is predefined. The task is to understand the basis for the classification from this data set (called training or learning set) which is achieved by using various classification algorithms. This information is then used to classify future samples into one of the predefined classes. Several classification algorithms have been introduced in the past. For application to microarray data two of the most frequently used methods are the nearest shrunken centroids method and support vector machines (SVMs). Both are explained in more detail below.

2.5.4.1 Hierarchical clustering

In hierarchical clustering the data is not partitioned in a single step. Instead, a series of partitions takes place which may run from a single cluster containing all objects to *N* clusters each containing a single object or vice versa. Hierarchical clustering belongs to the so-called agglomerative methods which proceed by series of fusions of the *N* objects into groups.

Given a set of *N* objects to be clustered as well as a *N* x *N* distance matrix, hierarchical clustering proceeds as follows:

1. Start by assigning each object to its own cluster, resulting in *N* clusters.
2. Find the closest pair of clusters and merge them into a single cluster.
3. Compute distances between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size *N*.

There are a number of methods that can be used to compute the new distances in step 3. Most commonly used are single-linkage, complete-linkage and average-linkage clustering.

Single-linkage clustering

In single-linkage clustering the distance between two clusters A and B is defined as the shortest distance from any member of one cluster to any member of the other cluster.

$$Distance\ (A, B) = min_{\mathrm{x}\in\mathrm{A},\mathrm{y}\in\mathrm{B}}\{d(x, y)\}$$

Complete-linkage clustering

In complete-linkage clustering, the distance between two clusters A and B is defined as the greatest distance from any member of one cluster to any member of the other cluster.

$$Distance\ (A, B) = max_{\mathrm{x}\in\mathrm{A},\mathrm{y}\in\mathrm{B}}\{d(x, y)\}$$

Average-linkage clustering

In average-linkage clustering, the distance between two clusters A and B is defined as the average distance from any member of one cluster to any member of the other cluster.

$$Distance\ (A, B) = \frac{\mathrm{T_A B}}{\mathrm{N_A} \times \mathrm{N_B}}$$

Where $\mathrm{T_A B}$ is the sum of all pairwise distances between cluster A and B and $\mathrm{N_A}$ and $\mathrm{N_B}$ are the sizes of clusters A and B, respectively.

For calculating the actual distance d of two objects in each of these methods, a number of different distance measures are commonly used, including the Euclidean distance and the Pearson correlation coefficient.

Euclidean distance

$$d = \sum_{i=1}^{n} \sqrt{(x_i - y_i)^2}$$

Pearson correlation coefficient

$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

Where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the average of the values in x and σ_x is the standard deviation of these values, similarly for $\bar{y}$ and σ_y.

2.5.4.2 Nearest shrunken centroid classification

Nearest shrunken centroid classification is a supervised classification method and is an enhancement of the nearest centroid classification method (Tibshirani, Hastie et al. 2002).

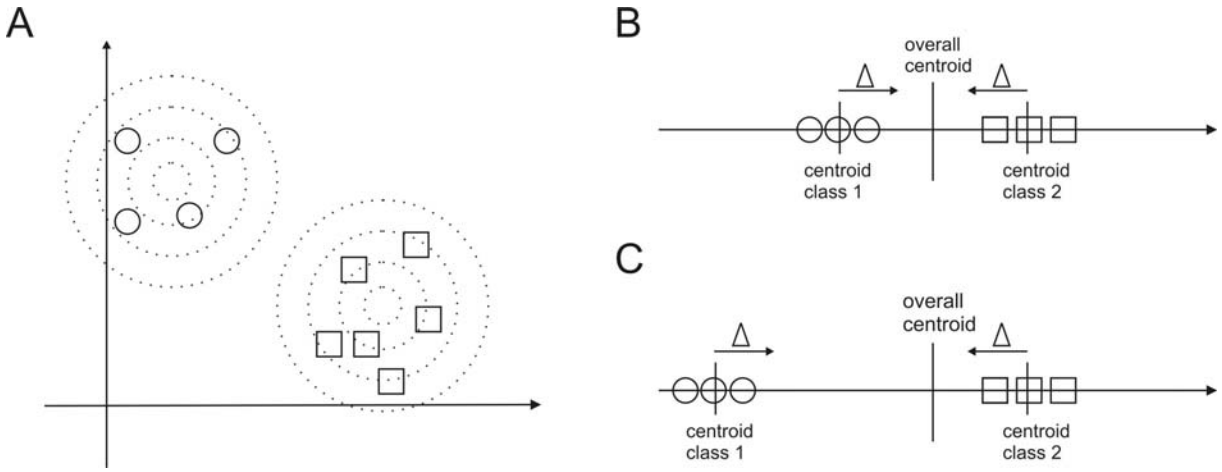


Figure 2.4 – Nearest shrunken centroid classification
(**A**) The nearest centroid method is a precursor method to nearest shrunken centroid classification and computes a standardized centroid for each class. This is the average gene expression for each gene in each class divided by the within-class standard deviation for that gene. In nearest shrunken centroid classification the class centroids are shrunk (by Δ) towards the overall centroid for all classes. A gene which is shrunk to zero for all classes, i.e. had similar expression values in al classes is eliminated from the analysis (**B**). A gene which is shrunk to zero for all classes except one is used for classification (**C**).

Briefly, the nearest centroid method computes a standardized centroid for each class. This is the average gene expression for each gene in each class divided by the within-class standard deviation for that gene (**Figure 2.4A**). The method then takes the gene expression profile of a new sample, and compares it to each of these class centroids. The class whose centroid that it is closest to, in squared distance, is the predicted class for that new sample. Nearest shrunken centroid classification makes one important modification to standard nearest centroid classification. For each gene, it shrinks each of the class centroids toward the overall centroid for all classes by an amount Δ. This shrinkage consists of moving the centroid towards zero by Δ, setting it equal to zero if it hits zero (**Figure 2.4B, C**). Genes that are shrunk to zero for all classes are eliminated from further analysis (**Figure 2.4B**). Alternatively, genes that are shrunk to zero for all classes except one are then characterizing that class by high or low expression (**Figure 2.4C**). After shrinking the centroids, the new sample is classified by the usual nearest centroid rule, but using the shrunken class centroids.

2.5.4.3 Support Vector Machines

A support vector machine (SVM) is not a real machine, but a mathematical method used in pattern recognition which has also been introduced as an approach for classification purposes (Vapnik 1998). A SVM divides a set of objects into classes, so that the area between the class borders is maximized (**Figure 2.5**). Since the SVM approach is a supervised classification approach, the starting point is objects for which the class affiliations are known. Each object is thereby represented by a vector.

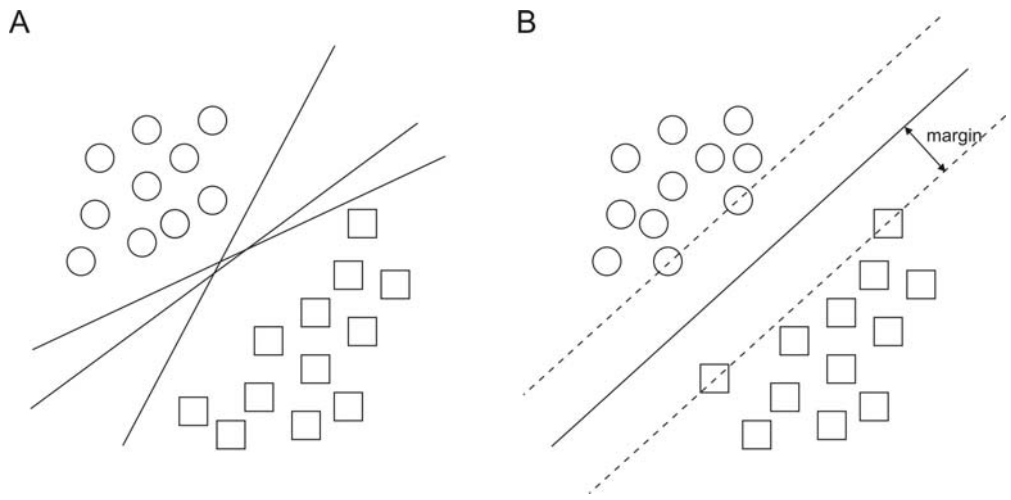


Figure 2.5 – Fitting a hyperplane
In supervised classification, objects are assigned to predefined classes, here, circles and squares. In an SVM approach, each object is represented by a vector in a multi-dimensional vector space. To separate the classes, many possible hyperplanes can be fit into the data (**A**). A SVM aproach constructs a hyperplane with the greatest area between the class borders by maximizing the margin between the vectors which are closest to the hyperplane and the hyperplane itself. These vectors are called support vectors (**B**).

The SVM now fits a hyperplane into this vector space which will separate the objects into two classes. Since there are many possibilities for such a hyperplane (**Figure 2.5A**), the SVM constructs a hyperplane which shows the greatest margin to the vectors which are closest to the hyperplane (**Figure 2.5B**). These vectors are called support vectors. The larger the margin, the better the classification of objects will be. A hyperplane cannot be bent; a separation is therefore only possible for objects which are linearly separable like in **Figure 2.5**. This is usually not the case in real-world applications (**Figure 2.6A**). In this case SVMs use the so-called kernel trick to still fit a hyperplane to the data.

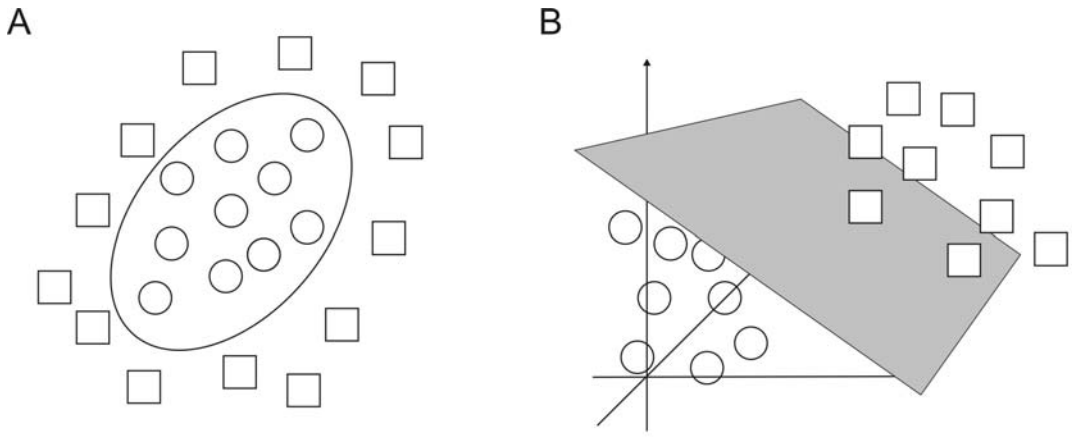


Figure 2.6 – Using the kernel trick to separate objects which are not linearly separable
The objects in real-world applications are usually not linearly separable (A). The SVM approach therefore uses the kernel trick to transform the vector space and the sample vectors to a higher dimensional space in which the objects are linearly separable. Here the hyperpane is constructed and sample vectors and vector space are transformed back to the original space.

The idea behind the kernel trick is to transform the vector space and also the sample vectors into a higher dimensional space. In this high-dimensional vector space the objects are linearly separable and the hyperplane can be constructed (**Figure 2.6B**). When transforming the vector space back to the lower dimensional space the hyperplane becomes a non-linear hyperplane which separates the objects into two classes.

2.6 The concept of RNA fingerprints

With the introduction of microarray technology, researchers have started to describe changes in gene expression between different samples (Schena, Shalon et al. 1995; DeRisi, Penland et al. 1996; Lockhart, Dong et al. 1996; Spellman, Sherlock et al. 1998). This research was performed in a rather

descriptive than analytical way and included the determination of differentially expressed genes and the grouping of genes based on their expression pattern across samples using unsupervised classification methods, as for example hierarchical clustering. When researchers started using clustering approaches on samples within an experiment (DeRisi, Penland et al. 1996; Lander 1996; Khan, Simon et al. 1998; Kononen, Bubendorf et al. 1998), the question arose whether there is a possibility to systematically classify diseases based on these transcriptional changes. In 1999, Golub and colleagues for the first time performed class prediction to classify new, unknown samples based on their distinct expression profiles. Using their proposed algorithm they developed a 50-gene predictor and accurately predicted all samples according to the patients' clinical diagnosis (Golub, Slonim et al. 1999). Reams of different algorithms have been introduced since then, all aiming on the sub-classification, prediction and diagnosis of different diseases (Vapnik 1998; Tibshirani, Hastie et al. 2002). Additionally, the type of tissue has been experimented with, starting from cell lines, biopsy material (Bhattacharjee, Richards et al. 2001; Pomeroy, Tamayo et al. 2002) to peripheral blood (Alizadeh, Eisen et al. 2000; Shipp, Ross et al. 2002; Valk, Verhaak et al. 2004).

In 2003, Joseph Nevins' lab introduced an elegant approach for prediction of oncogenic pathway activity in mouse tissue (Huang, Ishida et al. 2003). Huang and colleagues determined specific gene signatures for different oncogenes *in vitro* and applied these gene signatures to tumor samples to demonstrate the existence of these molecules *in vivo*. In 2006, Bild and colleagues adapted this approach to human cells (Bild, Yao et al. 2006). By transfecting normal human cells with single oncogenes followed by genome-wide transcriptional analysis they determined a specific gene signature for each of these oncogenes. Using descriptive and analytical bioinformatics, this gene signature was then applied to genome-wide transcriptional profiles of human malignancies. Here they clearly demonstrated that these oncogene-specific signatures can be recognized within the malignant cells.

We took up the described approach of *in vitro* generation of specific gene signatures and hypothesized that - in principle - this concept should be applicable to any other cell and factor that leads to transcriptional changes upon stimulation and signaling. In a research project concerned with the interrogation of immune inhibition within the tumor environment, we generated gene signatures for different inhibitory molecules. We termed the generated gene signatures RNA fingerprints of the interrogated molecules. These RNA fingerprints should then provide direct evidence whether the cells within a tumor environment are under the control of the interrogated molecules (Chemnitz, Eggle et al. 2007). The term RNA fingerprint can therefore be described as the transcriptional changes which are observed in response to a biological stimulus, in this case the response to an inhibitory molecule. Following this concept we hypothesized that, in general, transcriptional changes

which are observed in response to any molecular signal can be termed RNA fingerprints. These signals include activated oncogenic pathways by introducing the specific oncogene (Vapnik 1998; Huang, Ishida et al. 2003; Bild, Yao et al. 2006), the stimulus of an inhibitory molecule, as described above, or the response to a given disease. A RNA fingerprint of a given disease is therefore a specific image for this disease which can then be used for diagnostic and predictive purposes. Furthermore, pre-defined RNA fingerprints of different processes and functions can be derived from establishes databases, as for example the Gene Ontology (GO) consortium (Ashburner, Ball et al. 2000) or the KEGG database (Kanehisa, Araki et al. 2008). All mentioned examples of RNA fingerprints will be introduced in the course of this thesis.

3 Material and Methods

3.1 Sample collection and isolation of cells

Generation of *in vitro* RNA fingerprints

For the generation of *in vitro* RNA fingerprints of TGFβ and PD-1, blood samples were collected from healthy blood donors after informed written consent was obtained in accordance with the Declaration of Helsinki. $CD4^+$ T cells were isolated by negative selection as described previously (Chemnitz, Driesen et al. 2006). Lymph node specimens of 9 patients with classic Hodgkin lymphoma (HL), 9 patients with FL, and 9 patients with reactive lymph node reaction (RLN) of different causes were included. This study was performed within the framework of the German Hodgkin Study Group. When possible, samples were taken at primary diagnosis. Also included were 3 samples with aberrant diagnosis: 1 patient with T-cell–rich B-cell lymphoma (B-NHL); 1 with lymphocyte-predominant HL (LPHL), but with tumor-free tissue in the removed lymph node; and 1 with HL, with histologically proven follicular lymphoma in prior medical history. $CD4^+$ T cells from lymph node specimens were isolated by mechanical homogenization of the specimen and subsequently purified by positive selection on ice using magnetic cell sorting columns (Miltenyi Biotech, Bergisch Gladbach, Germany) according to the manufacturer's instructions. All samples were taken after informed consent following approval by the Ethik Kommission of the University of Cologne, Cologne, Germany.

Generation of a disease specific RNA fingerprint

In the prevalent cohort 2.5 ml blood was drawn directly into PAXgene vials providing stabilization of the gene expression profile. Samples were rested over night at room temperature and then stored at -80°C until further preparation. In the incident cohort snap frozen PBMC enriched blood (~ 300 µl) from the EPIC study was used for RNA extraction. Blood samples were directly thawed in 5 ml of TRI Reagent BD (Molecular Research Center, Inc, USA).

Pre-defined RNA fingerprints

Blood samples from healthy blood donors were collected after written informed consent had been obtained. $CD4^+$ T cells were isolated from blood samples by using a RosetteSep $CD4^+$ enrichment kit (StemCell Technologies); purity was >90% as determined by flow cytometry.

3.2 RNA preparation and microarray hybridization

For all microarray experiments using the Illumina BeadChip technology, RNA was isolated according to the manufacturer´s protocol with subsequent column purification using the RNeasy MinElute Cleanup Kit (Qiagen, Hilden, Germany). Total RNA from PAXgene samples was prepared according to the manufacturer´s recommendations including an optional DNAse digestion step. cDNA and biotin-labeled cRNA synthesis was generated from 100 ng total RNA using the Illumina® TotalPrep™ RNA Amplification Kit (Applied Biosystems, Darmstadt, Germany). cRNA (1.5 µg) was hybridized to Human-6 Expression BeadChips V1 and V2 (Illumina, San Diego, CA) and scanned on Illumina BeadStation 500x. For microarray experiments using the Affymetrix GeneChip technology, RNA isolation, quantification and target preparation was performed according to standard protocols for small samples and cRNA was hybridized to HG-U133A arrays.

3.3 Statistical and bioinformatic data analysis

Raw data collection for Illumina BeadChip and Affymetrix HG-U133A arrays was performed using Illumina© BeadStudio software or Affymetrix MAS5.0 software. Further statistical and bioinformatic analyses were performed using R language (R Development Core Team 2007) and packages from the Bioconductor project (Gentleman, Carey et al. 2004). For normalization of data from the two platforms we used quantile and invariant set normalizations implemented in the affy package. Differentially expressed genes were selected using a fold change/p-value filter with the following criteria: fold change ≥ 2, absolute difference in signal intensity between group means ≥ 100 and p-value ≤ 0.05. Hierarchical cluster analysis was performed using the hcluster package. Before clustering the data was $\log_2$ transformed. Distances of the samples were calculated using a correlation coefficient (correlation similarity metric) and clusters were formed by taking the average of each cluster (average linkage). PCA analysis was performed using the pcurve package in R. When visualizing PCA results the first three principal components (coordinates) were plotted in 3-dimensional space. For supervised classification, the pamr package which uses the shrunken centroid method and the e1071 package for support vector machine classification are used.

Part II: Different concepts of RNA fingerprints

4 *In-vitro* generated RNA fingerprints

In this chapter the term RNA fingerprint is introduced in the context of an unsolved question in immunology, the immune inhibition within the tumor environment in humans. We generate gene signatures for different T cell inhibitory molecules *in vitro* and introduce these signatures as RNA fingerprints of the molecules. These fingerprints are then applied to gene expression profiles of human cancers to directly determine the *in vivo* impact of the interrogated molecules on tumor infiltrating T cells.

4.1 Biological motivation

A hallmark of various human malignancies is the expression of immunoinhibitory factors within the tumor microenvironment. There is indirect evidence based on *in vitro* experiments that tumor-infiltrating T cells in human malignancies are suppressed by such factors. Still, direct evidence of the influence of individual inhibitory factors on immune cells in human cancer *in vivo* is lacking. To address this question we used Hodgkin's lymphoma (HL) to determine whether HL cells are under the control of a particular inhibitory factor. HL qualifies as a model since its histopathological characteristics are thought to be mostly due to the effects of a wide variety of cytokines, including TGFβ or membrane bound receptors like PD-1. These cytokines are suspected to contribute to immune evasion of tumor cells.

4.2 Results

The approach of generating *in vitro* RNA fingerprints from gene expression profiles should – for the first time – provide direct evidence whether a particular cell is indeed under the control of a particular inhibitory factor *in vivo*. We established specific TGFβ and PD-1 RNA fingerprints in human $CD4^+$ T cells and applied these RNA fingerprints to transcriptional profiles of $CD4^+$ T cells isolated from HL lymph nodes. To determine whether the influence of TGFβ on $CD4^+$ T cells is specific for HL or can also be detected in other lymphomas, we also applied the fingerprints to $CD4^+$ T cells

originated from follicular lymphoma (FL), thereby providing direct evidence that these inhibitory factors are clearly signaling in T cells infiltrating HL but not FL.

Quantification of the inhibitory effect of TGFβ and PD-1 on human CD4$^+$ T cells *in vitro*

To directly determine the in vivo impact of inhibitory cytokines such as TGFβ or inhibitory surface receptors such as PD-1 on tumor infiltrating T cells we postulated that the factor-dependent transcriptional regulation assessed on a genome-wide scale should be comparable in T cells directly isolated from tumor tissue and T cells exposed to TGFβ or PD-1 in vitro. Prior to assessment of transcriptional changes as a consequence of stimulation with TGFβ or PD-1 we established the functional impact of both factors on highly purified CD4$^+$ T cells derived from healthy donors. The impact of TGFβ resp. PD-1 was assessed in context of T cell receptor mediated activation since it has been previously shown that T cells in vivo would be exposed to inhibitory factors in the context of antigen recognition within the tumor microenvironment (Poppema 1996; Poppema, Potters et al. 1998; Lin, Medeiros et al. 2004).

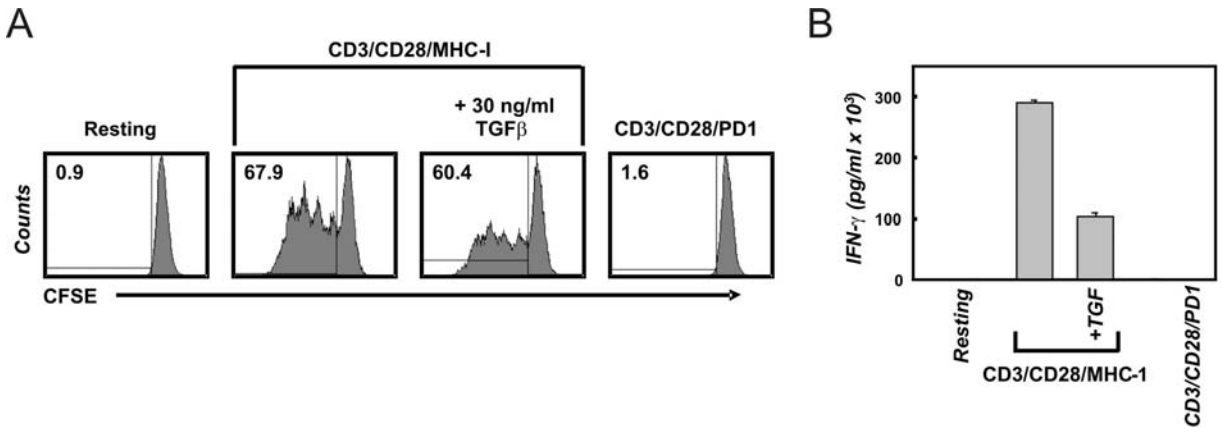


Figure 4.1 – Inhibition of T cell proliferation and IFN-γ secretion by TGFβ and PD-1
(**A**) Freshly isolated primary human CD4$^+$ T cells were labeled with CFSE and left unstimulated or were stimulated with the indicated magnetic beads (artificial antigen presenting cells, CD3/CD28/MHC-I resp. CD3/CD28/PD1) in the absence or presence of 30 ng/ml TGFβ. After 4 days CFSE dilution was analyzed by flow cytometry. The overall percentage of dividing cells is displayed inside the corresponding dot plot. (**B**) CD4$^+$ T cells were stimulated as above. After four days of incubation the concentration of IFN-γ was determined using flow cytometric bead assays. The presented data is representative for at least 3 independent experiments, error bars in *B* represent triplicates of one representative experiment.

The CD4$^+$ T cells were labeled with 5,6-Carboxyfluorescin-Diacetat-Succinimidyl-Ester (CFSE) and subsequently stimulated with aAPC (CD3/CD28/MHC-I) with or without TGFβ or aAPC coated with CD3/CD28/PD-1 for up to 96 hours (**Figure 4.1** and **Appendix C – Supplementary Methods,** shown here is the 96 hour time point). As expected, stimulation of primary CD4$^+$ T cells with

CD3/CD28/MHC-I resulted in robust T cell expansion and cytokine secretion. Addition of TGFβ to the cultures reduced T cell proliferation, albeit this effect was not as dramatic as that induced by PD-1 stimulation, which completely inhibited T cell proliferation (**Figure 4.1A**). In contrast, IFN-γ secretion was clearly decreased by both TGFβ and PD-1 (**Figure 4.1B**).

TGFβ and PD-1 RNA-fingerprints in CD4$^+$ T cells from healthy donors

For establishing the TGFβ resp. PD-1 fingerprints, CD4$^+$ T cells from 4 donors were either left unstimulated (resting cells) or stimulated with CD3/CD28/MHC-1 (activated cells) with or without addition of TGFβ (TGFβ-treated cells) or were stimulated with aAPC coated with CD3/CD28/PD-1 (PD-1 treated cells). To filter genes regulated under direct influence of TGFβ or PD-1 we analyzed transcriptional changes in two different ways: In a first step a) resting cells *vs.* activated cells and b) resting cells *vs.* TGFβ- resp. PD1-treated cells were compared. Genes specifically regulated under the influence of TGFβ resp. PD-1 were determined using set theory supported by Venn diagrams as previously described (Chemnitz, Driesen et al. 2006).

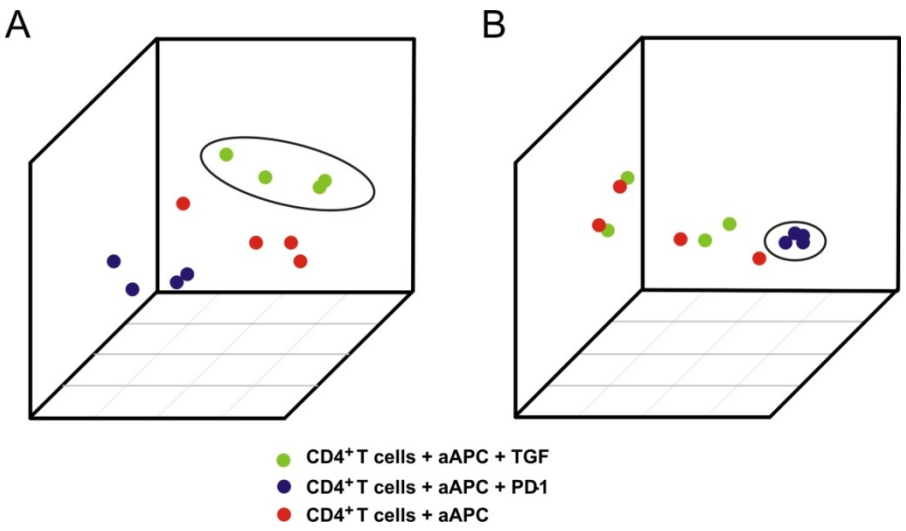


Figure 4.2 – Generalization of TGFβ and PD-1 genomic fingerprints
Principal components analysis (PCA) using the RNA-fingerprints of TGFβ or PD-1 was performed to distinguish samples treated with TGFβ or PD-1 from the respective control samples. The first three principal components derived from (**A**) the TGFβ and (**B**) the PD-1 fingerprints are plotted. Samples stimulated with magnetic beads coated with CD3/CD28/MHC-I are depicted in red, samples treated with CD3/CD28/PD-1 in blue and samples treated with CD3/CD28/MHC-I in the presence of TGFβ in green, respectively.

In a second step we compared expression profiles of activated cells *vs.* TGFβ- resp. PD1-treated cells. We defined the union of lists from step one and two as direct impact of either factor on the CD4$^+$ T

cell transcriptional profile and thus as the RNA-fingerprints of TGFβ and PD-1 signaling in T cells (112 respectively 37 genes, see (Chemnitz, Eggle et al. 2007), **Supplemental Table S1, S2**).
When plotting the first three principal components derived from the two signatures, samples treated with TGFβ (**Figure 4.2A**) or PD-1 (**Figure 4.2B**) were accurately distinguished from the other samples. This clear separation of samples treated with TGFβ or PD-1 from the remaining samples documents the particular impact of TGFβ and PD-1 on $CD4^+$ T cells and therefore provides the rationale for defining these signatures as RNA-fingerprints.

$CD4^+$ T cells in Hodgkin's lymphoma differ from T cells of reactive lymph nodes

To first assess overall differences between $CD4^+$ T cells derived from HL and FL versus RLN we performed a descriptive bioinformatics analysis. $CD4^+$ T cells from RLN were used as a control reflecting the characteristics of healthy T cells to the closest point possible. FL was used as a second malignancy to determine disease specific differences. For this analysis we used expression profiles of 5 samples from RLN patients, 4 samples from HL patients and 3 samples from FL patients derived from the Affymetrix HG-U133A microarray. Genes were defined as differentially regulated if FC >2 or FC <-2, p-value < 0.05 and difference in sample means > 100. In total we found 108 differentially expressed genes between $CD4^+$ T cells derived from HL and RLN samples (42 up-, 66 down-regulated) and 144 differentially expressed genes between $CD4^+$ T cells derived from FL resp. RLN samples (144 down-regulated) (see (Chemnitz, Eggle et al. 2007), **Supplemental Table S3**). Interestingly, when comparing for T cell activation induced genes no significant differences between the patient groups were observed (data not shown). To link differential expression of genes to biological processes, we postulated that it is possible to apply the RNA-fingerprints we established for TGFβ and PD-1 in our in vitro system to answer the question, whether such inhibitory mechanisms play a role in HL *in vivo*.

$CD4^+$ T cells in Hodgkin's lymphoma harbor the TGFβ fingerprint

To separate distinct sample groups based on different biologies several approaches including unsupervised as well as supervised approaches have been developed. If TGFβ indeed acts on $CD4^+$ T cells in HL, it should be possible to correctly separate T cells isolated from HL from $CD4^+$ T cells isolated from RLN within the gene space of the TGFβ fingerprint established *in vitro*. We therefore applied a total of 4 independent approaches, namely (i) hierarchical clustering, (ii) principal component analysis (PCA), (iii) classification based on nearest shrunken centroids (PAM), and (iv) support vector machines (SVM). We first performed this analysis on the Affymetrix platform on a

subgroup of patients, namely 5 samples from RLN patients, 4 samples from HL patients and 3 samples from FL patients. By applying hierarchical clustering using the TGFβ fingerprint HL and RLN were separated into two distinct clusters (**Figure 4.3A**). Correct separation was still achieved when using less stringent filter criteria for generating the TGFβ fingerprint (data not shown).

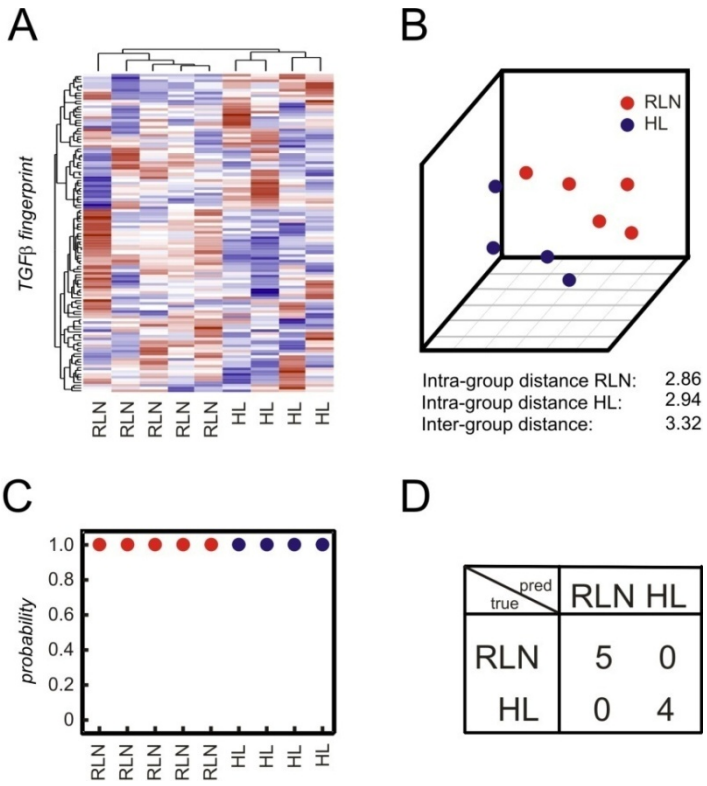


Figure 4.3 – HL samples are separated from RLN samples on the basis of TGFβ regulated genes
The RNA-fingerprint of TGFβ was used to separate transcriptional profiles of HL from RLN. **(A)** Hierarchical cluster analysis using average linkage and correlation distance metric. (**B**) Result of principal components analysis (PCA) with the first three principal components is shown. (**C**) Supervised classification using PAM; for each sample the posterior probability, i.e. the percentage of certainty of a correct class prediction is plotted. (**D**) Supervised classification using SVMs. A fourfold table comparing the predicted class labels to the actual class labels is depicted.

In contrast, when applying other gene sets established as biologically defined RNA-fingerprints including the predictive gene signatures established by Bild et al (Bild, Yao et al. 2006), HL and RLN samples were not correctly separated. This analysis included fingerprints associated with transcriptional changes following activation of Ras, Myc, E2F3, Src, b-catenin, EGF, VEGF, or NFκB respectively fingerprints associated with T-cell activation, cell cycle activity, apoptosis, inflammatory response, or chemokine activity (data not shown). These findings further support the specificity of the TGFβ fingerprint within the HL samples. As a second unsupervised approach we applied PCA. When plotting the first three principal components HL and RLN samples were again separated using

the TGFβ fingerprint. This was further supported by a larger inter-group distance (between HL and RLN) compared to the respective intra-group distances (**Figure 4.3B**). To more formally assess the existence of a TGFβ fingerprint signature in HL we applied leave-one-out cross validation based on PAM and SVMs. PAM analysis predicted HL respectively RLN cases with a 100% accuracy and posterior probability based on the genes within the TGFβ fingerprint (**Figure 4.3C**). Using the SVM approach, again, a 100% accuracy was achieved (**Figure 4.3D**). So far, assessment of differential transcriptional regulation in $CD4^+$ T cells from either HL or RLN based on specific RNA-fingerprints indicated that TGFβ is an important component of the HL environment leading to signaling events in $CD4^+$ T cells infiltrating the tumor site.

PD-1 signaling is also prominent in T cells derived from Hodgkin's lymphoma

The same four bioinformatics approaches were used to determine whether genes of the PD-1 fingerprint were also harbored in HL-derived $CD4^+$ T cells.

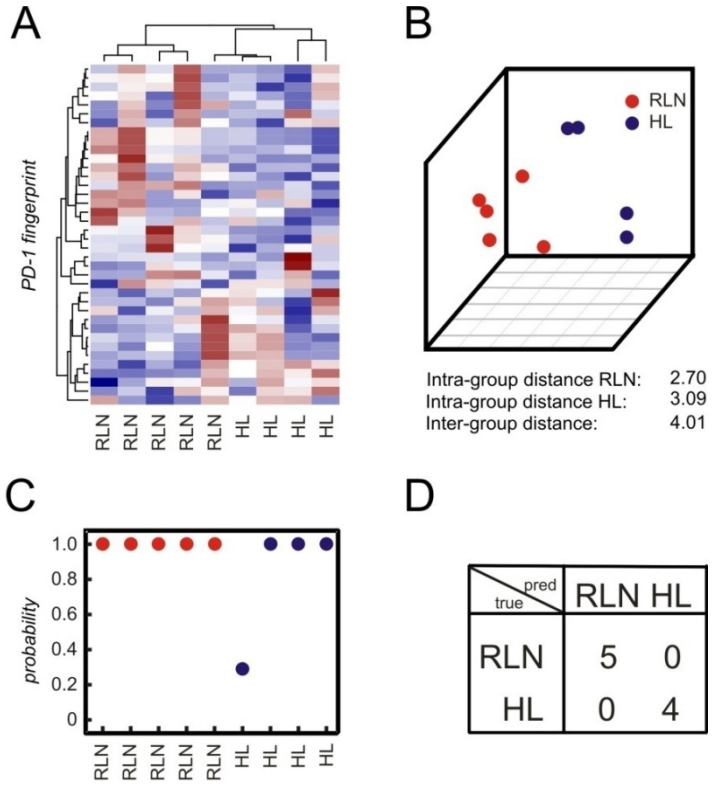


Figure 4.4 – HL samples are separated from RLN samples on the basis of PD-1 regulated genes
The RNA-fingerprint of PD-1 was used to separate transcriptional profiles of HL from RLN. (**A**) Hierarchical cluster analysis using average linkage and correlation distance metric. (**B**) Result of principal components analysis (PCA) with the first three principal components is shown. (**C**) Supervised classification using PAM. (**D**) Supervised classification using SVMs.

As depicted in **Figure 4.4A**, HL and RLN samples were correctly separated when applying hierarchical clustering based on the PD-1 fingerprint. Similarly, applying PCA led to a correct separation of HL and RLN samples which was also supported by a larger inter-group distance (**Figure 4.4B**). When applying PAM, one sample was always falsely predicted and the posterior probability never reached 100% for all samples (**Figure 4.4C**). Using SVM though, the prediction accuracy was 100% based on the PD-1 fingerprint (**Figure 4.4D**). Taken together, the results indicate PD-1 to be a further important factor in the HL environment.

RNA fingerprints reveal no impact of TGFβ or PD-1 on CD4$^+$ T cells in follicular lymphoma

To determine whether the influence of TGFβ on CD4$^+$ T cells is specific for HL or can also be detected in other lymphomas, we analyzed CD4$^+$ T cells derived from patients with FL. We first assessed the influence of TGFβ by applying the TGFβ fingerprint.

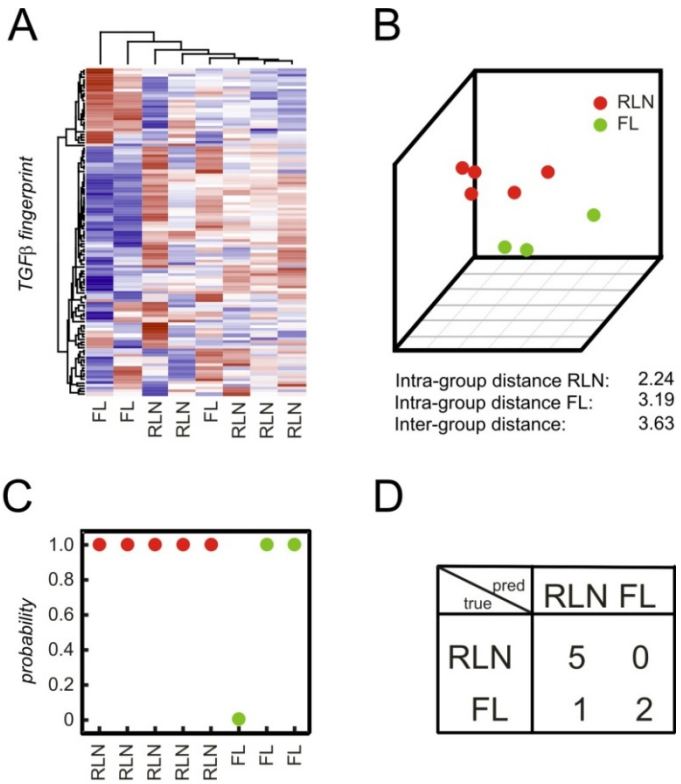


Figure 4.5 – FL samples are not separated from RLN samples on the basis of TGFβ regulated genes
The RNA fingerprint of TGFβ was used to separate transcriptional profiles of FL from RLN. (**A**) Hierarchical cluster analysis using average linkage and correlation distance metric. (**B**) Result of principal components analysis (PCA) with the first three principal components is shown. (**C**) Supervised classification using PAM. (**D**) Supervised classification using SVMs.

When using hierarchical clustering, FL samples never correctly separated from RLN samples (**Figure 4.5A**). Similarly, the supervised approaches showed no correct prediction (**Figure 4.5C**, **D**). Only when applying PCA, FL and RLN samples were correctly separated and the inter-group variance was larger than the intra-group distances (**Figure 4.5B**). Similarly, none of the above mentioned fingerprints (e.g. Ras, Myc) correctly separated FL from RLN samples indicating that none of these pathways play a major role in $CD4^+$ T cells derived from FL tissue.

Next we applied the PD-1 fingerprint, however, none of the four tests achieved a correct separation of FL and RLN samples (**Figure 4.6**). We conclude from these analyses that TGFβ and PD-1 do not induce major transcriptional changes in $CD4^+$ T cells from FL.

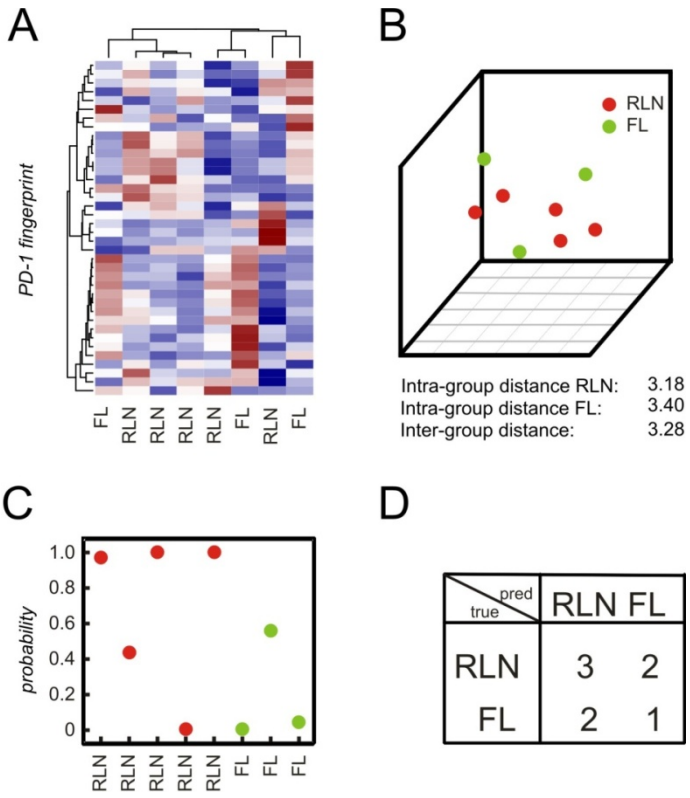


Figure 4.6 – FL samples are not separated from RLN samples on the basis of PD-1 regulated genes
The RNA fingerprint of PD-1 was used to separate transcriptional profiles of FL from RLN. (**A**) Hierarchical cluster analysis using average linkage and correlation distance metric. (**B**) Result of principal components analysis (PCA) with the first three principal components is shown. (**C**) Supervised classification using PAM. (**D**) Supervised classification using SVMs.

Validation of the method using additional patient samples and a different array platform

To validate our method and to show the independency of the results from the microarray platform we used the Illumina© BeadChip platform for further analysis. Here we analyzed 5 patients with HL, 6

patients with FL and 4 patients with RLN. Additionally we included three samples with aberrant diagnosis to further specify our approach: one patient with T-cell-rich B-cell-Lymphoma (B-NHL), one patient with Lymphocyte-Predominant HL (LPHL), however tumor free tissue in the removed lymph node specimen and one patient with HL, however histologically proven FL in the prior medical history. First, we tested the TGFβ fingerprint. As depicted in **Figure 4.7A** the TGFβ fingerprint correctly separates the HL samples from the RLN samples with only one HL sample falsely allocated to RLN.

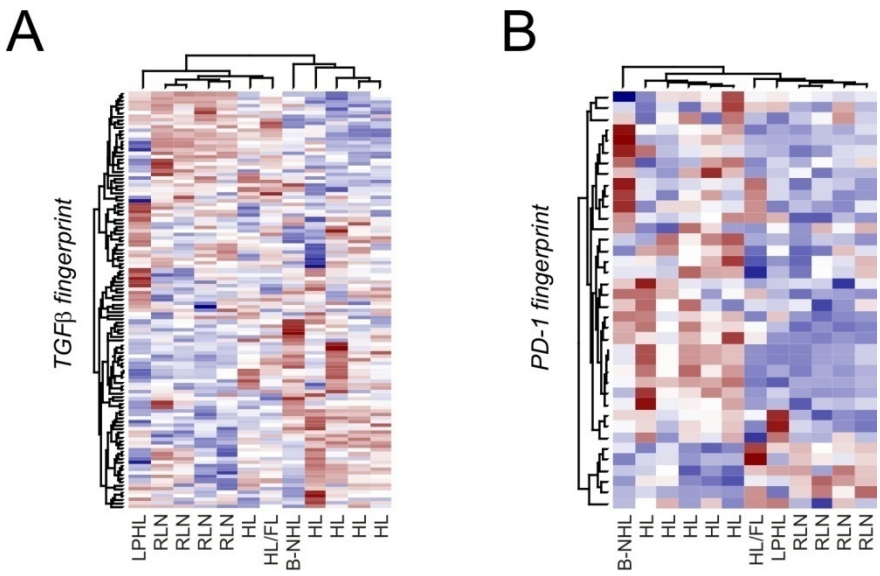


Figure 4.7 – Validation of the results using additional patient samples and a second array platform
$CD4^+$ T cells were isolated from lymph nodes of five patients with HL, 4 patients with RLN and 6 patients with FL. Three samples with aberrant diagnosis are labeled as follows: B-NHL: T-cell-rich B-cell-lymphoma, LPHL: Lymphocyte-Predominant Hodgkin's Lymphoma, HL/FL: Hodgkin Lymphoma with premedical history of Follicular Lymphoma. The RNA-fingerprints of (**A**) TGFβ and (**B**) PD-1 were used to differentiate HL and RLN samples using hierarchical clustering.

Interestingly, T cells derived from a tumor-free lymph node of a patient with LPHL clustered together with the RLN samples, suggesting that TGFβ mediated signaling events are restricted to the tumor in HL. Similarly, T cells from the patient with prior history of FL were more closely related to T cells from RLN samples. The results of the PCA analysis mirrored the hierarchical clustering. Moreover, both supervised approaches resulted in a significant classification of the different samples, highlighting the impact of TGFβ on $CD4^+$ T cells in HL (data not shown).

When using the PD-1 fingerprint, HL and RLN samples were correctly separated. **Figure 4.7B** displays the results of hierarchical clustering. The results of the PCA analysis and both supervised methods confirmed the separation and correct classification of the different samples (data not

shown). Taken together, even when using a different array platform, both, TGFβ and PD-1 fingerprints separate HL from RLN samples. This result gives further evidence for the impact of both, TGFβ and PD-1 on $CD4^+$ T cells in the tumor microenvironment of HL.

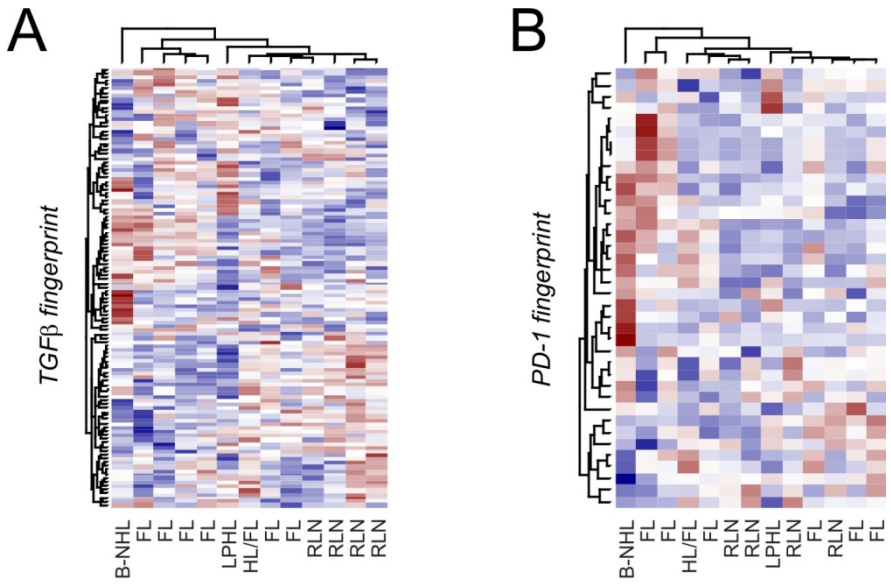


Figure 4.8 – Validation of the results using additional patient samples and a second array platform
$CD4^+$ T cells were isolated from lymph nodes of five patients with HL, 4 patients with RLN and 6 patients with FL. Three samples with aberrant diagnosis are labeled as follows: B-NHL: T-cell-rich B-cell-lymphoma, LPHL: Lymphocyte-Predominant Hodgkin's Lymphoma, HL/FL: Hodgkin Lymphoma with premedical history of Follicular Lymphoma. The RNA-fingerprints of (**A**) TGFβ and (**B**) PD-1 were used to differentiate FL and RLN samples using hierarchical clustering.

When analyzing $CD4^+$ T cells from FL patients, FL and RLN samples were not separated by hierarchical clustering using either the TGFβ (**Figure 4.8A**) or the PD-1 (**Figure 4.8B**) fingerprint. Also, PCA and both supervised methods failed to classify the samples accordingly (data not shown) thereby supporting the specificity of both factors towards HL.

Cross-platform analysis further supports the impact of TGFβ and PD-1 on $CD4^+$ T cells in HL but not in FL

To analyze all samples together irrespective of array platform used we applied an approach for cross-platform analysis introduced by Warnat et *al.* (Warnat, Eils et al. 2005). Due to the quantitative nature of the method, hierarchical clustering was not a useful tool for analyzing data derived from different array platforms since it regularly separates samples based on technology used rather than biology. In contrast, supervised approaches can be performed on data derived by cross-platform

analysis. As shown in **Figure 4.9A**, PAM analysis predicted HL respectively RLN samples with 79% accuracy and high posterior probabilities based on the genes within the TGFβ fingerprint.

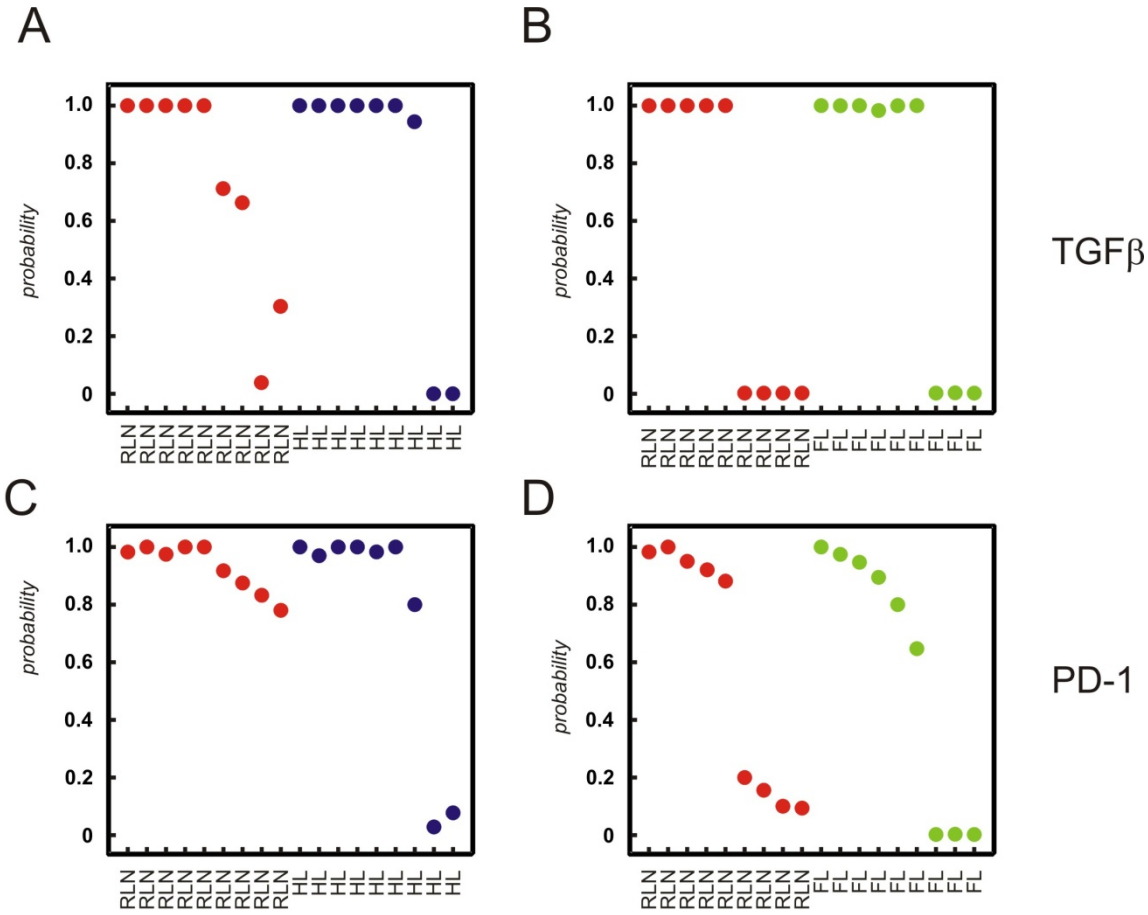


Figure 4.9 – Combined analysis of all samples irrespective of array platform used
Supervised classification (PAM) using the data derived from cross-platform analysis; 9 samples from HL patients, 9 samples from FL patients and 9 samples from patients with a RLN were used for analysis. For each sample the posterior probability, i.e. the percentage of certainty of a correct class prediction is plotted. The TGFβ fingerprint was used to classify (**A**) HL and RLN samples or (**B**) FL and RLN samples, respectively. The PD-1 fingerprint was used to classify (**C**) HL and RLN samples or (**D**) resp. FL and RLN samples.

Using the PD-1 fingerprint we derived a total accuracy of 89% (**Figure 4.9C**). On the opposite, when classifying FL and RLN samples, the overall prediction accuracy was only 53% for both the gene spaces of the TGFβ and the PD-1 fingerprints (**Figure 4.9B**, **D**)**.** Again, we verified the specificity of the fingerprints, this time by analyzing 335 biologically defined gene spaces (terms defined by Gene Ontology; GO Terms) chosen based on size of the respective GO Terms (including 50-100 genes). Less than 9% of these gene spaces derived a correct classification of HL versus RLN samples respectively FL versus RLN samples (data not shown). This data further strengthens the hypothesis for both, TGFβ and PD-1 to play a role in inhibiting $CD4^+$ T cells specifically in HL but not FL.

4.3 Discussion and further research options

We have adapted the recently introduced approach of determining predictive gene signatures *in vitro* (Bild, Yao et al. 2006) to derive a concept of RNA fingerprints. Applying this concept to the problem of immune inhibition within a tumor, we provided direct evidence that RNA fingerprints of T cells derived from healthy individuals activated in the presence of inhibitory cytokines such as TGFβ or inhibitory receptors like PD-1 can be used to directly determine, whether T cells isolated from diseased tissue are indeed under the influence of these inhibitory factors *in vivo*. Moreover, we showed that both, TGFβ and PD-1 have distinct impact on $CD4^+$ T cells in HL but not in FL.

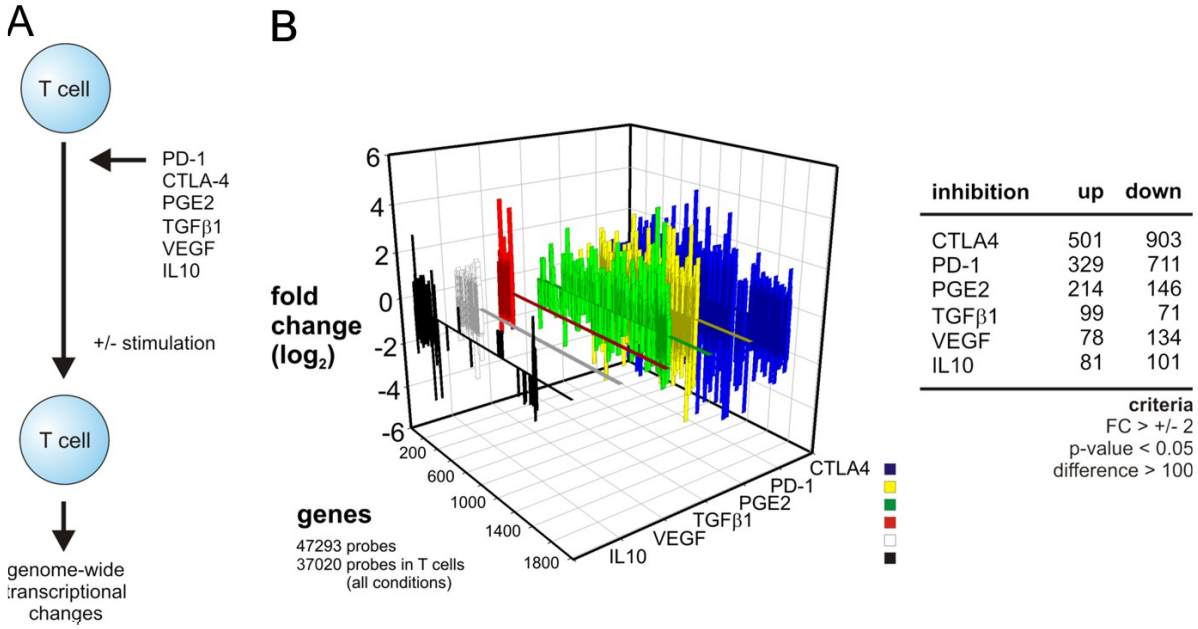


Figure 4.10 – Defining RNA fingerprints of different inhibitory molecules
Fingerprints for CTLA4, PD-1, PGE_2, TGFβ, VEGF and IL10 were generated to quantify their influence on $CD4^+$ T cells. Experiments were performed as described above (**A**). The generated fingerprints were plotted next to each other to visually find differences and similarities (**B**)

An extension of this study has already been initiated which deals with the generation of RNA fingerprints of different inhibitory molecules and the quantification of their influence on $CD4^+$ T cells. In addition to the already established fingerprints of PD-1 and TGFβ we have generated fingerprints for CTLA4, PGE_2 (Chemnitz, Driesen et al. 2006), VEGF and IL10. For each of these inhibitory molecules, in vitro experiments were performed as described above (**Figure 4.10A**) and fingerprints were established accordingly. Genes which comprise the distinct RNA fingerprints were plotted next to each other to visually find differences and similarities (**Figure 4.10B**). It is apparent that the fingerprints of CTLA4, PD-1 and PGE_2 share common features which are quite distinct from

fingerprints derived from TGFβ, VEGF and IL10. These common and distinct features resemble the separation of the molecules in molecules with high inhibitory effect (CTLA4, PD-1 and PGE_2) and lower inhibitory effect (TGFβ, VEGF and IL10). Right now methods are implemented which numerically quantify the inhibitory effect of the different molecules. Additionally we are working on distinguishing the RNA fingerprints of both the molecules with high inhibitory effect and molecules with lower inhibitory effect to get specific inhibitory features of these molecules. This will hopefully lead to further functional characterization of the signaling pathways these molecules are involved in.

5 Disease specific RNA fingerprints

5.1 Motivation

As demonstrated in the previous chapter, we were able to use a biologically defined *in vitro* fingerprint to predict an actual *in vivo* situation. This chapter introduces the concept of a disease specific fingerprint. Here, the biologically defined RNA fingerprint of lung cancer was used to predict the occurrence of lung cancer prior to clinical manifestation.

5.2 Biological motivation

Lung cancer is the most frequent cause of cancer related death in the western world. Prognosis has remained disastrous during the last decades with a median overall 2 year survival rate of only 10% (Mountain 1997). This is mainly due to late detection of the disease and therefore the development of efficient tools for early detection thus represents the most promising strategy to improve prognosis of lung cancer (Mulshine 2003). Numerous screening approaches have been tested over the last decades including chest X-ray, spiral computed tomography (CT) and identification of oncogene mutations, microsatellite losses and epigenetic changes (Swensen 2003; Bremnes, Sirera et al. 2005; Ganti and Mulshine 2005; Swensen, Jett et al. 2005). None of these approaches was a real breakthrough for early detection of lung cancer. Additionally, several limitations, as for example the high costs and radiation exposure for spiral CT are apparent. Very recently Spira and colleagues analyzed histologically normal large-airway epithelial cells obtained at biopsy from smokers with suspicion of lung cancer (Spira, Beane et al. 2007). Using gene expression profiling they compared smokers with and without subsequent diagnosed lung cancer and identified a 80-gene biomarker that distinguished these two groups. In two validation studies they demonstrated the predictive ability of the biomarker with a mean sensitivity of 80%. These findings indicate that gene expression in cytologically normal large-airway epithelial cells can serve as a lung cancer biomarker which can be used for early detection of lung cancer. Up to now several studies revealed the potential gene expression profiling to establish predictive marker or signatures for diagnosis and prognosis of different diseases in peripheral blood (Alizadeh, Eisen et al. 2000; Shipp, Ross et al. 2002; Valk, Verhaak et al. 2004). We postulated that gene expression profiles of peripheral blood samples derived from patients with manifest lung cancer can be used to develop a RNA fingerprint of lung cancer. We further postulated that these transcriptional changes which are associated with lung

cancer might be an early event in lung cancer development and might therefore be suitable as marker for early detection.

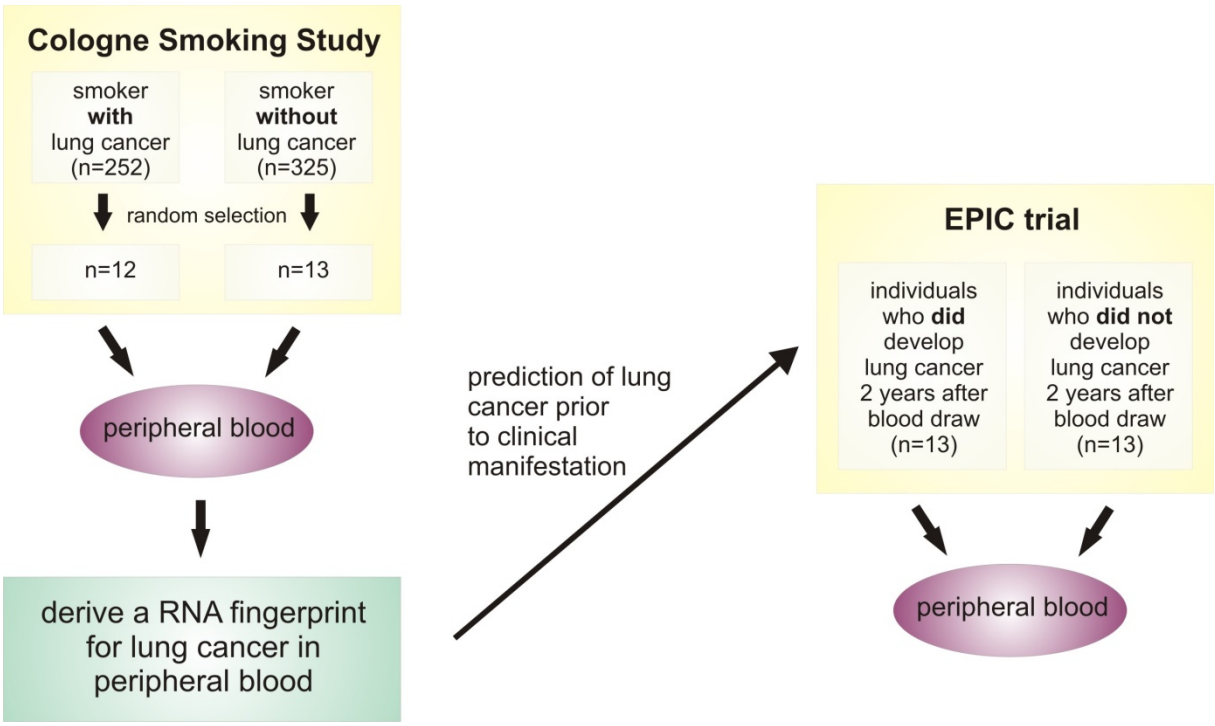


Figure 5.1 – Strategy for predicting lung cancer prior to clinical manifestation
Peripheral blood derived from both, diagnosed lung cancer patients and patients without diagnosed lung cancer were used to obtain a RNA fingerprint predictive for lung cancer. This RNA fingerprint was then used to predict lung cancer prior to clinical manifestation.

We therefore used expression profiles of peripheral blood cells derived from both, diagnosed lung cancer patients and patients without diagnosed lung cancer to obtain a RNA fingerprint of lung cancer. Phrasing this procedure in terms of the concept introduced in the preceding chapter, the generation of this lung cancer associated RNA fingerprint was now performed *in vivo*. We then asked the question whether patients who will develop lung cancer in the future already exhibit this RNA fingerprint in their peripheral blood cells. We therefore tested the predictive nature of the RNA fingerprint and applied it to gene expression profiles of patients with developing lung cancer but prior to clinical manifestation. **Figure 5.1** depicts an overview of the strategy.

5.3 Results

To determine a lung cancer associated RNA fingerprint in peripheral blood, samples from 13 active smokers with clinically manifest lung cancer comprising 11 patients with non-small cell lung cancer (NSCLC) and 2 patients with small-cell lung cancer (SCLC) as well as 11 control samples from cancer free smokers were studied (LC cohort). In a first step we determined differentially expressed genes between SCLC, NSCLC and controls in the LC cohort using an ANOVA based filter (p-value < 0,003). The ANOVA based filter was used to derive all genes which are differentially expressed in at least one of the three groups. 151 genes satisfied the criteria and were referred to as the RNA fingerprint of clinically manifest lung cancer. This fingerprint was then used for further analysis. When performing hierarchical clustering bases on the RNA fingerprint a clear separation of the three different groups (NSCLC, SCLC and controls) was demonstrated within this data set (**Figure 5.2**). The predictive ability of the fingerprint was further demonstrated by a leave-one-out cross-validation within the data set using the complete RNA fingerprint as a predictor (according to Chapter 4.2).

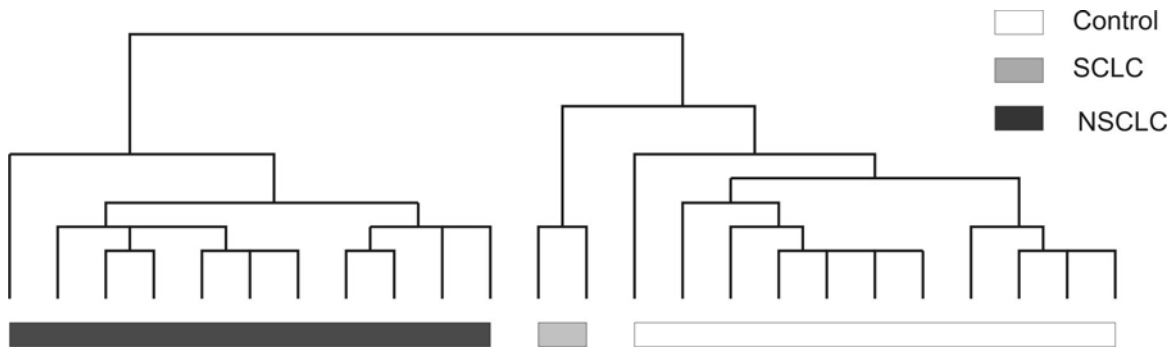


Figure 5.2 – Hierarchical clustering distinguishes SCLC patients from NSCLC patients and controls
Differentially expressed genes between SCLC, NSCLC and controls in the LC group using an ANOVA based filter (p-value < 0,003) and used these genes for hierarchical clustering.

To address the question whether patients who will develop lung cancer in the future already exhibit the lung cancer associated RNA fingerprint in their peripheral blood cells we applied the RNA fingerprint to the Heidelberg cohort of the EPIC trial (EPIC cohort). Within EPIC, data and biological material from about 500.000 people from 10 European countries have been collected; the Heidelberg cohort includes 25543 probands. Within this cohort, 14 actively smoking individuals had developed either NSCLC (n=8) or SCLC (n=6) within 24 months (median 14 months) post sample asservation. As controls, a group of active smokers (n=16), who had not developed lung cancer within 10 years post sample asservation was chosen based on matching of gender, age and smoking

behavior. Within this data set quality control was especially crucial, as samples were stored as whole blood and frozen without RNA stabilization.

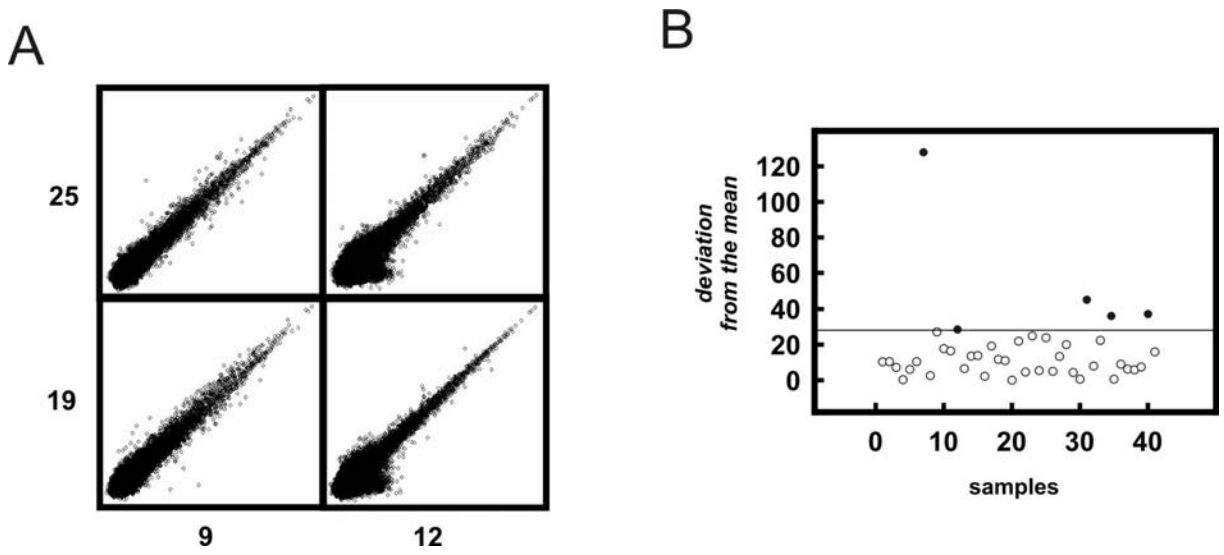


Figure 5.3 – Quality control of samples from the EPIC cohort
For quality control pairwise scatterplots were used in a first step (**A**). Then we calculated the deviation from the overall mean for each of the arrays to detect outlier arrays (**B**).

Using the common quality measures described in Chapter 2.5.1 we determined 5 outlier arrays where we detected systematic differences of low expression values (**Figure 5.3B**, shown are two exemplified plots). Additionally we calculated the absolute deviation of each array from the overall median. In short, the median expression value for each array was calculated. Next the median of these medians (overall, median) was taken and the deviation of each array median from the overall median was determined. When plotting the deviations (**Figure 5.3A**) 4 samples (7, 31, 34, 40) clearly showed large deviations from the median (37.2 to 127.9) and were removed from further analysis. Although sample 12 did not show a very large deviation from the median, we still removed the sample from further analysis due to the noticeable differences of low expression values (**Figure 5.3B**). The remaining samples (7 NSCLC samples, 5 SCLC samples and 13 controls) were used for analysis. When performing hierarchical clustering based on the established lung cancer associated RNA fingerprint we derived a separation of cases and controls; only 6 of the 25 samples were misgrouped (**Figure 5.4**). To demonstrate the significance of this finding, we performed a permutation analysis. We randomly assigned classes to samples in the LC cohort, identified differentially expressed genes between NSCLC, SCLC and controls (p-value < 0.003) and used them to perform hierarchical clustering in the EPIC cohort. From 1000 random permutations, only 3 resulted in a separation of cases and controls with less than 6 misclassifications. This corresponds to a p-value of 0.003.

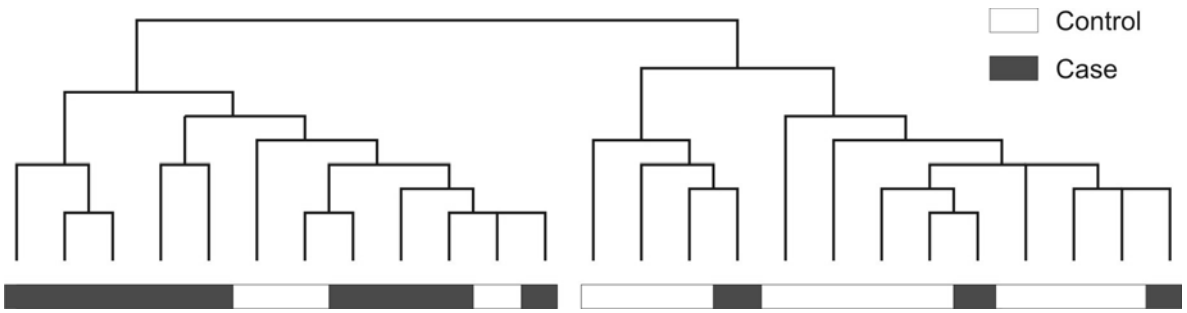


Figure 5.4 – Hierarchical clustering of samples from the EPIC cohort
Differentially expressed identified in the LC cohort were used for hierarchical clustering of samples from the EPIC cohort.

In a next step we performed prediction of developing lung cancer in the EPIC cohort. We therefore used the RNA fingerprint established in the LC cohort and built a predictor based on the K-nearest neighbor algorithm and validated it using leave one out cross validation (Gene Pattern, Boston USA). This analysis resulted in a 65 feature predictor which was subsequently used to predict the samples from the EPIC cohort. 5 of the 25 samples were not correctly predicted which corresponds to a 80% prediction accuracy (**Figure 5.5**).

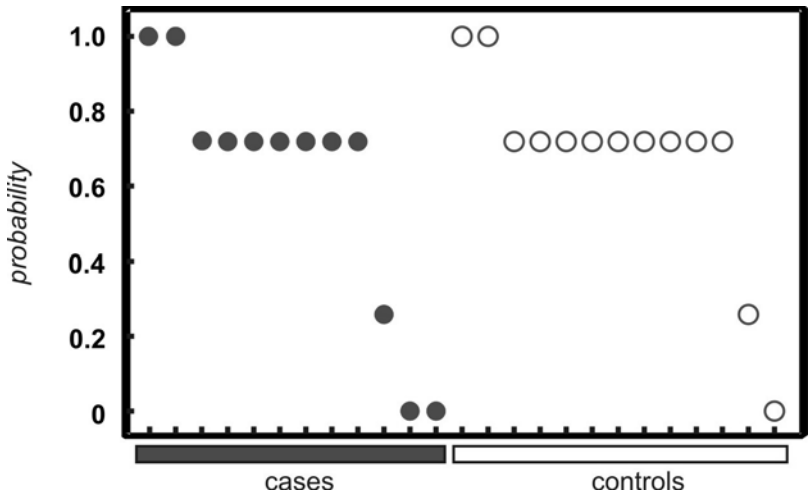


Figure 5.5 – Prediction of cases and controls from the EPIC cohort
The list of differentially expressed genes from the LC cohort was used to build a predictor based on the K-nearest neighbor algorithm and validated using leave one out cross validation (Gene Pattern, Boston USA). A 65 feature predictor was used to predict the samples from the EPIC cohort. Depicted is the prediction probability for each sample.

5.4 Discussion and further research options

In this study we demonstrated differential expression of several genes between patients with lung cancer and controls and generated a RNA fingerprint predictive of lung cancer. Several other studies detected similar differences in the absolute amount of RNA as well as the expression of single genes e.g. c-met, hnRNP B1, hTERT ERCC1, XPD, RAI) (Bremnes, Sirera et al. 2005). We further show that this RNA fingerprint can be used to detect individuals developing lung cancer prior to clinical manifestation. The ability to define a group at high risk of developing lung cancer in smoking adults with an easily applicable blood test may be very valuable to define candidates for further early detection techniques such as spiral CT.

This study was a prospective study on the prediction of lung cancer prior to clinical manifestation. To substantiate the findings within this study, the study has to be repeated within a larger setting, composing at least 200 patients with lung cancer and 200 controls for the creation of the RNA fingerprint. Additionally, the EPIC cohort has to be enlarged to analyze another 200-400 samples. We have already started to at least substantiate the predictive ability of the RNA fingerprint on a validation cohort composed of another 37 samples including 22 patients with manifest lung cancer.

6 Pre-defined RNA fingerprints

In the preceding chapters we have defined RNA fingerprints based on our own experiments. Additionally to self-defined RNA fingerprints one could consider the data stored in biological databases as pre-defined RNA fingerprints. Biological databases include information about genes belonging to special pathways or groups of genes with similar functions. These groups of genes could definitely be considered as RNA fingerprints of pathways or functions. Applying these pre-defined RNA fingerprints to a microarray experiment and searching for patterns of these fingerprints in the data is called gene-class testing. Here, I will introduce the idea of gene class testing and will provide a new method implementing this approach.

6.1 The idea of gene-class testing

Several methods have been applied to gene expression data in order to detect changes in expression between different subsets of samples (Golub, Slonim et al. 1999; Tusher, Tibshirani et al. 2001). Since most of these methods result in a list of differentially expressed genes the main challenge of biologists lies in interpreting these long lists to extract biological meaning. Therefore, gene-class testing (GCT) has been suggested as a powerful strategy to assess genome-wide gene expression data. In GCT gene classes are determined by mapping genes to biological pathways using annotations provided by the Gene Ontology (GO) Consortium (Ashburner, Ball et al. 2000). The GO Consortium provides controlled vocabularies which model Biological Process, Molecular Function and Cellular Component. Using a hierarchical tree structure gene products are annotated to one or more GO nodes according to their function. The GO Consortium is the most widely accepted gene annotation database and is updated in a daily manner. Many different tools have been introduced which analyze gene expression data using a GCT approach. While all of them share the common approach of searching for GO Terms enriched in a list of differentially expressed genes, they use different statistical models including hypergeometric, binomial, χ^2 and Fisher's exact test. An overview of 14 different tools can be found in (Khatri and Draghici 2005).

6.2 The "gold standard": Gene set enrichment analysis (GSEA)

More recently a new method, gene set enrichment analysis (GSEA) has been introduced (Mootha, Lindgren et al. 2003; Patti, Butte et al. 2003; Petersen, Dufour et al. 2004; Subramanian, Tamayo et al. 2005). GSEA follows a completely different approach and overcomes some of the major disadvantages of earlier tools. While other tools are dependent on a list of genes which have been called significant at an arbitrarily predefined threshold (usually a p-value) and therefore lose information from genes not satisfying the exclusion criterion, GSEA considers all genes in a dataset, irrespective of any arbitrary threshold. Also GSEA does not use a predefined statistical model, but uses an enrichment score which is statistically assessed using a permutation analysis. **Figure 6.1** shows an overview of the GSEA algorithm (Subramanian, Tamayo et al. 2005).

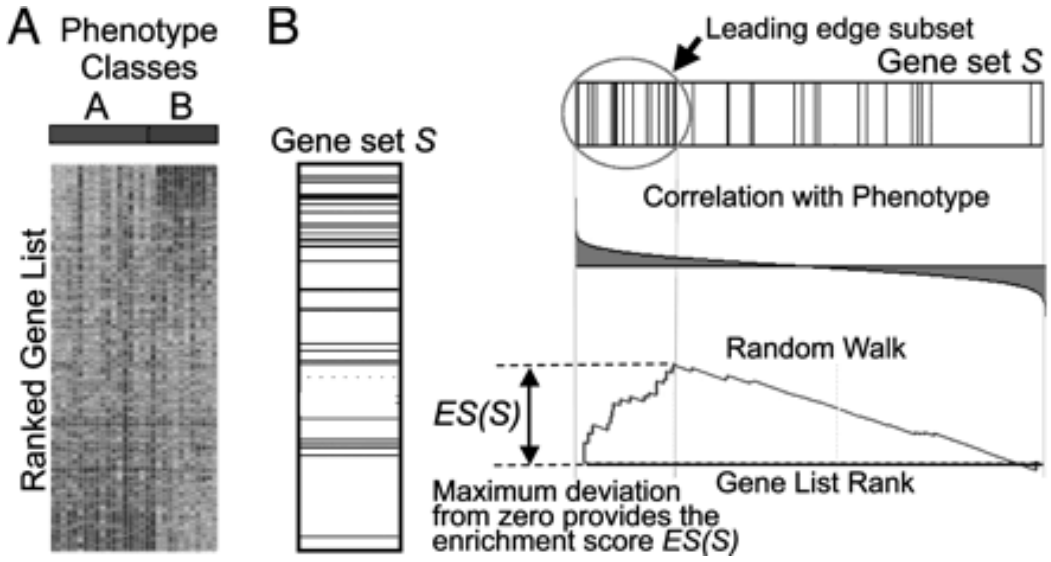


Figure 6.1 – GSEA overview
(**A**) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (**B**) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset (taken from (Subramanian, Tamayo et al. 2005)).

In short, genes are ranked based on the correlation between their expression and the class distinction by using any suitable metric, for example a test statistic, a signal-to-noise ratio or a fold change (**Figure 6.1A**). Then, for a given set of genes (e.g., genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same GO category) an Enrichment Score (ES) is calculated which reflects the degree to which a set of genes is overrepresented at the extremes (top or bottom) of the entire ranked list (**Figure 6.1B**). The significance level of ES is estimated by a permutation procedure in which the phenotypes of the samples are permuted. Lastly,

the estimated significance level is adjusted for multiple hypothesis testing (Subramanian, Tamayo et al. 2005).

6.3 A new approach: GOAna

Up to now GSEA is the "gold standard" of GCT tools, but albeit successfully applied, it suffers from some important disadvantages. One of the main drawbacks of this approach is that prior knowledge is needed to analyze a data set of interest. Computational issues (in the R version) keep the researcher from performing an unbiased approach, i.e. gene sets have to be filtered beforehand and will be biased towards previous knowledge and hypotheses. Also, the approach still depends on a ranking criterion which introduces a further bias towards genes which show e.g. a high signal to noise ratio or fold-change, respectively. Another disadvantage is the fact that GSEA is restricted to the analysis of two subgroups and cannot be extended to more than two groups. We therefore implemented a new algorithm called GOAna. GOAna, a Gene Ontology analysis tool assesses contributions of gene spaces, here GO Terms, to changes in gene expression between subgroups of an experiment. GOAna follows the simple approach of calculating distances of subgroups within predefined gene spaces and assesses their significance by sample permutation.

6.3.1 The algorithm

GOAna is based on the following algorithm (**Figure 6.2**):

1. Define gene spaces

Based on Gene Ontology (GO) classifications different gene spaces are determined in a four-step procedure:

1. Step: Retrieve GO IDs restricted to the specified category “Biological process”, “Molecular Function” or “Cellular Component”.
2. Step: Extract GO IDs represented on the array in use.
3. Step: Exclude probe sets from the GO IDs which are absent in more than 50% of the samples.
4. Step: Filter out gene spaces which include fewer than 5 probe sets**.**

2. Calculate distance between subgroups of the data

Let *n* be the number of genes within a pre-defined gene space, *k* the number of subgroups in the data. For each gene within a pre-defined gene space the mean expression value is calculated for each

subgroup. The resulting *n* x *k* matrix is transposed and the pairwise Euclidean distances between the subgroups of the data are calculated.

3. Assess significance of respective distance

Significance of the calculated distance, i.e. contribution of a respected gene space to changes in gene expression is assessed by a permutation analysis. Group assignments of the samples are permuted followed by a recalculation of the Euclidean distance (1000 times). Corresponding p-values are determined as the fraction of iterations where the distance obtained from the permuted groups is greater than the distance in the original data. The result is a list of GO IDs with an associated p-value.

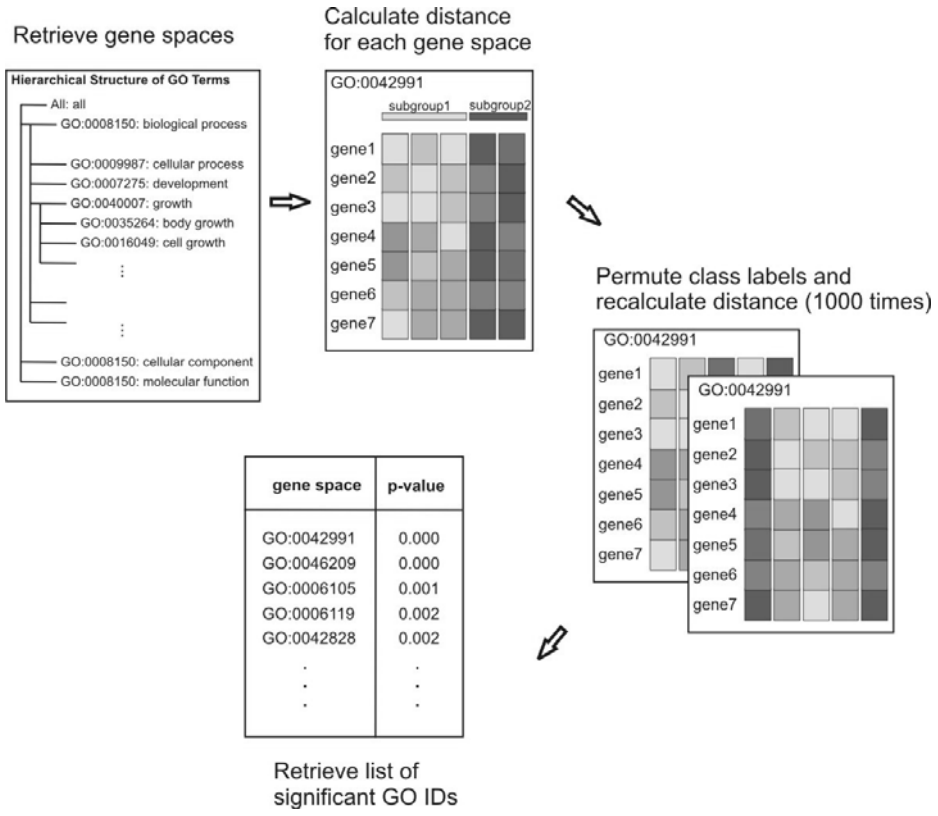


Figure 6.2 – GOAna overview
After retrieval of gene spaces, Euclidean distances between subgroups are calculated and significance is assessed by permutation and recalculation of distance (1000 times).

6.3.2 Proof of concept

To best possible compare GOAna and GSEA, we used the Diabetes data set published by Mootha et al. (Mootha, Lindgren et al. 2003). It consists of 43 skeletal muscle biopsy samples, 17 with normal

glucose tolerance (NGT), 8 with impaired glucose tolerance (IGT) and 12 with Type 2 diabetes mellitus (DM2). Normalized data was derived from the supplementary information website. Additionally the 149 manually curated gene sets analyzed in (Mootha, Lindgren et al. 2003) were retrieved through the same website. In their paper the authors identified a set of genes, called the OXPHOS gene set involved in oxidative phosphorylation whose expression is coordinately decreased in human diabetic muscle. We applied GOAna to the Diabetes data set and compared NGT patients to DM2 patients. In a first step, the GO category 'cellular component' was used. 376 GO IDs met the criteria of the algorithm (step 1) and were chosen for analysis. 24 of these were called significant with p-value < 0.05 (**Table 6.1**).

GOID	GOTerm	p-value
GO:0005730	Nucleolus	0.006
GO:0005759	mitochondrial matrix	0.007
GO:0005739	Mitochondrion	0.01
GO:0005952	cAMP-dependent protein kinase complex	0.012
GO:0043235	receptor complex	0.012
GO:0019717	Synaptosome	0.013
GO:0005750	respiratory chain complex III (sensu Eukaryota)	0.014
GO:0019866	inner membrane	0.014
GO:0042721	mitochondrial inner membrane protein insertion complex	0.014
GO:0005744	mitochondrial inner membrane pre-sequence translocase complex	0.015
GO:0008305	integrin complex	0.015
GO:0045275	respiratory chain complex III	0.016
GO:0005743	mitochondrial inner membrane	0.019
GO:0005746	mitochondrial electron transport chain	0.02
GO:0000776	Kinetochore	0.021
GO:0045202	Synapse	0.034
GO:0000775	chromosome, pericentric region	0.035
GO:0005593	FACIT collagen	0.036
GO:0005740	mitochondrial membrane	0.038
GO:0016469	proton-transporting two-sector ATPase complex	0.038
GO:0016323	basolateral plasma membrane	0.04
GO:0045263	proton-transporting ATP synthase complex, coupling factor F(o)	0.04
GO:0005892	nicotinic acetylcholine-gated receptor-channel complex	0.049
GO:0031090	organelle membrane	0.049

Table 6.1 – Significant GO IDs identified by GOAna
Running GOAna on 376 GO IDs derived from the category "cellular component", 24 were identified as significant with p-value < 0.05. Depicted is the GO ID, the description of the term and the calculated p-value.

It is apparent that most identified GO IDs are associated with the mitochondrion, indicating that the biological processes differing between DM2 patients and NGT patients are likely to occur at the

mitochondrion. When analyzing the data based on 1531 GO IDs from the category 'biological process' and 1138 GO IDs from the category 'molecular function', 96 respectively 69 GO IDs were called significant with p-value < 0.05.

GOID	GOTerm	p-value
GO:0007286	spermatid development	0
GO:0048468	cell development	0
GO:0048515	spermatid differentiation	0.001
GO:0051321	meiotic cell cycle	0.001
GO:0007283	spermatogenesis	0.001
GO:0019953	sexual reproduction	0.001
GO:0009303	rRNA transcription	0.001
GO:0006383	transcription from RNA polymerase III promoter	0.001
GO:0048232	male gamete generation	0.002
GO:0051327	M phase of meiotic cell cycle	0.002
GO:0007126	meiosis	0.002
…	…	…
GO:0042592	homeostasis	0.007
GO:0042990	regulation of transcription factor-nucleus import	0.008
GO:0019725	cell homeostasis	0.008
GO:0006105	succinate metabolism	0.008
GO:0006119	oxidative phosphorylation	0.009
GO:0042136	neurotransmitter biosynthesis	0.009
GO:0042345	regulation of NF-kappaB-nucleus import	0.01
GO:0042773	ATP synthesis coupled electron transport	0.01
GO:0042775	ATP synthesis coupled electron transport (sensu Eukaryota)	0.011
GO:0006118	electron transport	0.011
GO:0006753	nucleoside phosphate metabolism	0.012
GO:0050954	sensory perception of mechanical stimulus	0.013
GO:0006538	glutamate catabolism	0.013
GO:0006626	protein-mitochondrial targeting	0.014
GO:0046328	regulation of JNK cascade	0.015
GO:0000279	M phase	0.015
GO:0006536	glutamate metabolism	0.015
GO:0007007	inner mitochondrial membrane organization and biogenesis	0.016
GO:0045039	mitochondrial inner membrane protein import	0.017
GO:0007006	mitochondrial membrane organization and biogenesis	0.017
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	0.018

Table 6.2 – Excerpt of significant GO IDs retrieved by investigation of "biological processes"
1531 GO IDs from the category 'biological process' and 1138 GO IDs from the category 'molecular function' were analyzed. 96 respectively 69 GO IDs were called significant with p-value < 0.05. Depicted is the GO ID, the description of the term and the calculated p-value.

Within these results, we were able to recall oxidative phosphorylation, ATP metabolism and electron transport. **Table 6.2** shows an excerpt of the obtained GO IDs. However, these processes and functions did not achieve the most significant p-values within the analysis. The GO IDs identified as most significant by GOAna were all associated with the process of spermatogenesis in the context of cell development.

6.3.3 Discussion

The new algorithm introduced here, GOAna, implements a gene-class testing approach. This very simple approach makes it possible to perform an unbiased analysis bases on all branches of GO. When comparing the result obtained by GOAna with the one obtained by GSEA, the overall result, namely mitochondrial processes including electron transport and oxidative phosphorylation is identical. Indeed, genes included in the OXPHOS gene set identified by Mootha et al. are involved in oxidative phosphorylation which is associated with the mitochondrion. The OXPHOS genes identified by Mootha et al. comprise 114 hand-curated genes. When comparing this gene set with the GO ID 'Mitochondrion' 106 of 114 are included in both, the OXPHOS set and the GO ID. The detailed results though, are slightly different between GOAna and GSEA. Whereas GSEA identifies oxidative phosphorylation as the most significant biological process, GOAna identifies processes involved in spermatogenesis and cell development as most significant. The discrepancy between these results can be mainly explained by the different approaches taken. While GOAna was carried out using all processes included in GO, Mootha and colleagues used 149 hand-curated gene sets for their analysis. The process of spermatogenesis was not included in these gene sets. Therefore Mootha and colleagues were not even able to identify these processes. Indeed, when restricting the analysis to the 149 gene sets used by Mootha and colleagues, we were able to recall the same results as GSEA (data not shown). Therefore, we propose GOAna as an easy-to-use gene-class testing approach for unbiased analysis of microarray experiments.

6.4 Application to T cell homeostasis

6.4.1 Biological Motivation

Next, we applied GOAna to an immunological question, namely the field of T cell homeostasis. Based on studies in knockout mice, several exogenous inhibitory factors such as TGFβ, IL-10, or CTLA-4 have been implicated as gate keepers of adaptive immune responses. Lack of these inhibitory molecules

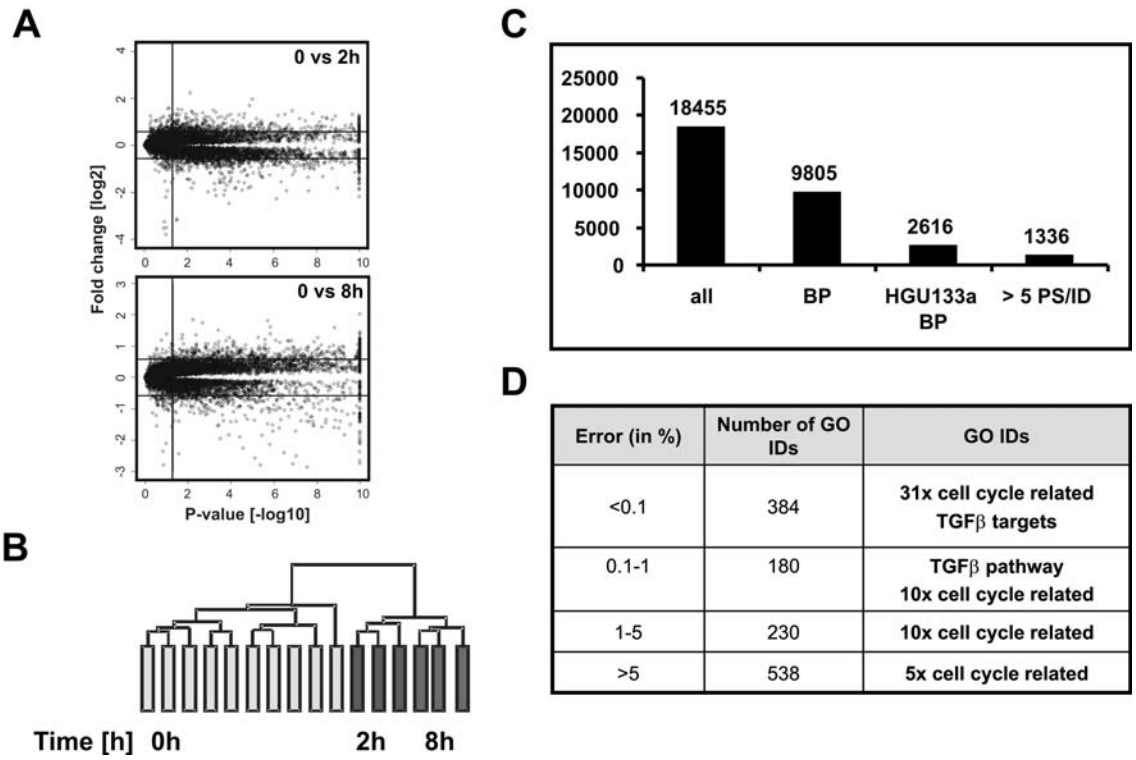
leads to massive inflammatory responses mainly mediated by activated T cells. In humans, the integration of these inhibitory signals for keeping T cells at a resting state is less well understood. It is tempting to speculate that the same factors involved in murine T cell homeostasis are also involved in human T cell homeostasis, especially because many of these factors have similar roles during induction of immunity. However, so far no experimental evidence exists supporting such a postulate. We therefore hypothesized that deprivation of resting human T cells of any exogenous signals should reverse intracellular signaling cascades actively keeping T cells at a resting state. We further postulated that such changes should certainly be recognizable on the genomic level. To this end, we interrogated genome-wide transcriptional changes of human mature $CD4^+$ T cells in response to deprivation of exogenous signals.

6.4.2 Results

To assess factors keeping T cells at a resting state, we exposed purified human $CD4^+$ T cells to an environment depleted of blood-derived soluble factors present in serum. Early genome-wide transcriptional changes were assessed using Affymetrix microarrays. Filtering based on fold changes (FC) and significance (variable probe sets, FC > 1.5 or FC < -1.5 and p-value < 0.05) revealed a high number of genes (878 genes, 443 up- and 435 down-regulated) with altered transcription after 2 h of serum deprivation in highly purified $CD4^+$ T cells. Changes of transcription even further increased at a later time point (910 genes at 8 h; 593 up- and 317 down-regulated) (**Figure 6.3A**). When performing hierarchical clustering based on all variable probe sets, time of serum withdrawal was the major factor separating the sample groups (**Figure 6.3B**).

Next, we were interested in determining which biological systems mainly contribute to these changes of gene expression. Therefore, we applied GOAna. In the first step we defined the set of gene spaces. Of the 18,455 currently known GO IDs, 9,805 comprise biological processes, 2,616 are present on the HGU133A array, but only 1,336 of them included at least 5 probe sets (**Figure 6.3C**). After the calculation of Euclidean distances between the three sample groups (time points t = 0, 2, and 8 h) and the significance analysis we identified 384 GO IDs to be affected on a significance level below 0.1%, 180 GO IDs between 0.1 and 1%, and 230 GO IDs between 1 and 5% after 8 h of serum deprivation (**Figure 6.3D**). When analyzing the most significant GO IDs, it became apparent that biological terms like cell cycle, cell growth, and transcription regulation were major contributors to differences in gene expression after serum deprivation. Surprisingly, 31 of 56 cell cycle-related GO IDs were affected on a significance level below 0.1% and only 5 cell cycle-related GO IDs did not reach the 0.05 significance level (error rate <5%). To identify signals that might account for these

changes in T cells after serum deprivation, we next searched for potential extrinsic signals upstream of cell cycle, cell growth, and transcription regulation.



D

Error (in %)	Number of GO IDs	GO IDs
<0.1	384	31x cell cycle related TGFβ targets
0.1-1	180	TGFβ pathway 10x cell cycle related
1-5	230	10x cell cycle related
>5	538	5x cell cycle related

Figure 6.3 – The TGFβ pathway is significantly changed by serum deprivation in CD4$^+$ T cells
(A) To visualize significant changes in gene expression, Volcano plots were used. The FC (log_2 FC) of gene expression was plotted against the negative p-value (log_{10}). All CD4$^+$ T cell samples assessed on the Affymetrix platform are included. Plotted are genes changed between t = 0 (n = 10) and t = 2 h (n = 3) as well as t = 0 and t = 8 h (n = 3). Genes significantly changed are defined by FC < -1.5 or FC > 1.5 and a p-value < 0.05 (see respective lines). **(B)** Hierarchical clustering of all T-cell experiments on the Affymetrix platform. Before clustering, genes were filtered using all variable probe sets (0.5 < SD/mean < 10). **(C)** All GO IDs (all) were filtered first on the category biological process (BP), next filtered on the presence on the HGU133a array (HGU133a/BP), and finally on those represented with at least five probe sets (>5 PS/ID). **(D)** GOAna revealed several significantly changed GO IDs (number of GO IDs) in CD4$^+$ T cells after 8 h of incubation in serum-free medium. The number of cell cycle-related GO IDs is given; highlighted on the right site are offsprings of the overall GO term cell cycle; *, TGFβ targets is a set of known TGFβ target genes. Significance levels are presented as error rates (in percent).

This analysis identified the TGFβ pathway to be the most significantly changed exogenous signaling cascade (error rate <1%). To corroborate the GO- based approach, a set of genes containing previously described TGFβ1 target genes (Siegel and Massague 2003; Renzoni, Abraham et al. 2004) was subjected to GO analysis. We postulated that these TGFβ target genes should again reveal significant changes in gene expression associated with serum deprivation. Indeed, this set of genes was even more significantly changed in human primary CD4$^+$ T cells (error rate <0.1%) (**Figure 6.3D**). To further evaluate the specificity of our results, GO IDs containing genes associated with immune

regulation were studied. Strikingly, none of these GO IDs reached a level of significance exceeding 1% (three GO IDs with an error rate between 1 and 5%, and eight GO IDs with an error rate <5%).

6.4.3 Discussion

In this study GOAna revealed that changes in TGFβ-related genes are major contributors to the overall transcriptional changes observed after serum deprivation in human $CD4^+$ T cells. GOAna was used as a starting point for further analyses to substantiate this finding (Classen, Zander et al. 2007). One way to show that indeed TGFβ was the exogenous factor keeping T cells at a resting state was the investigation of TGFβ target genes. As already described, the self-defined gene space of TGFβ target genes was called significant and therefore showed a contribution to the differences of different time points. When checking the signaling pathway of TGFβ using GenMAPP, several of the known TGFβ target genes induced upon TGFβ stimulation were shown to be under the permanent control of TGFβ in resting T cells. In the next step it was demonstrated that most of the known TGFβ target genes, which were identified as significantly regulated during serum deprivation were counterregulated after addition of TGFβ. Moreover, using this approach numerous novel TGFβ target genes were identified that are under the suppressive control of TGFβ. So far, these genes have not been recognized as TGFβ target genes in other cellular systems. Expression of these genes was up-regulated once TGFβ signaling was lost during serum deprivation and again suppressed upon TGFβ reconstitution. The other way to demonstrate constitutive TGFβ signaling in resting $CD4^+$ T cells was the interrogation of SMAD signaling which is an early event after the binding of TGFβ to its receptor complex. Indeed, immunofluorescence and Western Blot analysis demonstrated phosphorylated SMAD2 and SMAD3 in freshly isolated resting human $CD4^+$ T cells. Loss of transcriptional control by TGFβ should be accompanied by loss of SMAD phosphorylation which could be demonstrated in the paper. This phosphorylation could be restored by addition of either exogenous TGFβ or freshly isolated human serum (which contains TGFβ). Taken together, in our hands GOAna can be used in different experimental settings as a starting point for functional analyses.

7 The microarray experiment as a RNA fingerprint

Testing groups of genes or even whole pathways for differential expression was a huge advancement in the analysis of gene expression data. It enabled the researcher to detect a pattern of commonly regulated genes within a pathway when no differential expression was detected on a single gene level. However, there are still processes which cannot be identified by searching for differentially expressed genes or pathways, as for example processes which are not regulated on the transcriptional level. We therefore would like to take our GCT approach one step further and tackle the even more challenging question of whether the result of a gene expression analysis – we call it the RNA fingerprint of the microarray experiment – will give us a hint on what had happened upstream to the observed transcriptional changes.

7.1 Further development of GOAna

To achieve this goal, the GOAna algorithm was further developed to combine the already implemented GCT approach with a network construction approach.

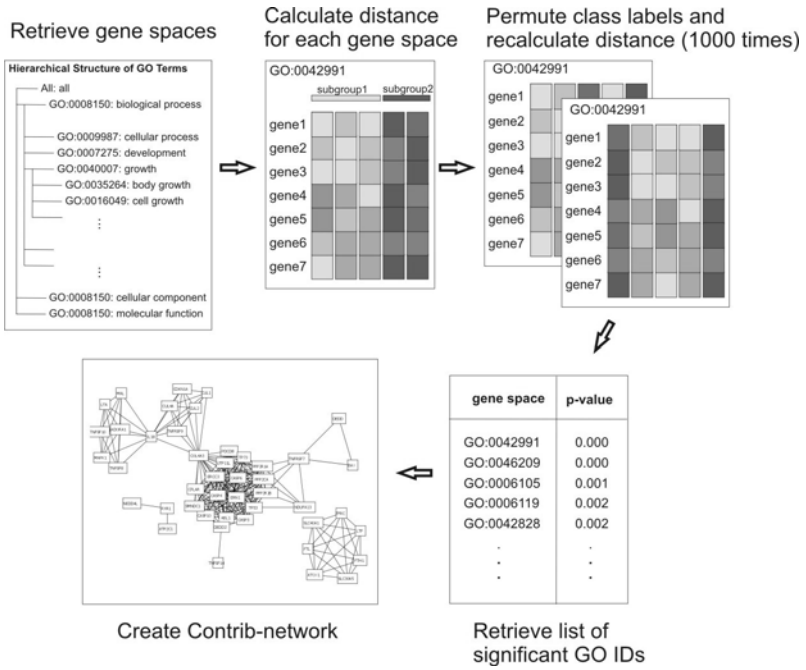


Figure 7.1 – Overview of the extended GOAna algorithm
After retrieval of gene spaces, Euclidean distances between subgroups are calculated and significance is assessed by permutation and recalculation of distance (1000 times). The hypothesis underlying the last step, the construction of the Contrib-network is depicted in more detail in **Figure 7.2**. In short, genes included in significant gene spaces are extracted and visualized using Cytoscape.

To date GOAna therefore includes two major analysis steps: (1) identification of biologically systems (defined by GO terms) affected by the experimental setting (original GOAna) and (2) identification of genes playing a central role in these biological systems. **Figure 7.1** briefly depicts the structure of the enhanced GOAna algorithm.

The first step of the algorithm has been introduced in Chapter 6.3 and needs no further introduction. The second step will be explained in detail here. After identification of the most significant gene spaces a network of contributing genes (Contrib-network) is constructed.

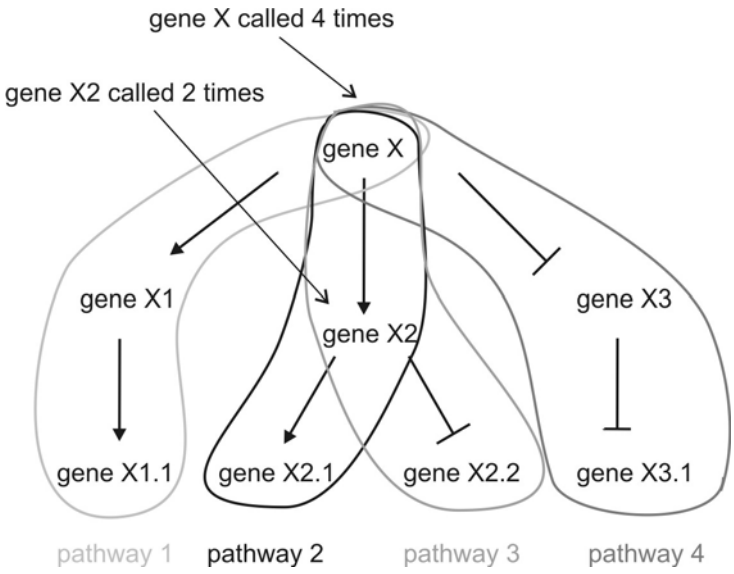


Figure 7.2 – Hypothesis for identification of key players
Graphical description of the major hypothesis for identifying upstream key players. Genes acting as major switches between gene spaces contributing to changes in gene expression between the analyzed subgroups will be included in several significant gene spaces and will therefore be extracted several times.

We hypothesized that genes in key regulatory upstream positions are likely to be involved in multiple biological processes (**Figure 7.2**) and would therefore appear several times in the list of significant GO IDs. To identify these genes, all significant GO IDs are used and the genes contained in these GOIDs are extracted. The network of contributing genes (contrib-network) is then constructed as follows:

Let $\mathrm{GOIDs}(gene_i)$ be the number of significant GO IDs which gene i is included in. Then $gene_i$ and $gene_j$ are connected by an edge if:

$$\mathrm{GOIDs}(gene_i) \geq 2 \;\&\&\; \mathrm{GOIDs}(gene_j) \geq 2 \qquad (1)$$

$$\mathrm{GOIDs}(gene_i) \cap \mathrm{GOIDs}(gene_j) \geq x \qquad (2)$$

Equation (1) makes sure that only genes which are represented by two or more GO IDs in the list of significant gene spaces are used for the construction of the contrib-network. In equation (2) the number of GOIDs x shared by $gene_i$ and $gene_j$ is specified. In an iterative process increasing x, genes which appear most often in the list of significant GO IDs can be determined.

7.2 Application to T cell inhibitors

To demonstrate the utility of the algorithm, we applied it to an unsolved biological question in T cell biology, the unraveling of detailed signaling mechanisms following inhibition of T cells. PGE_2 has diverse effects on $CD4^+$ T cells which lead to inhibition of T cell activation. Recently an interference of PGE_2 at an early step of T cell receptor signaling was suggested (Chemnitz, Driesen et al. 2006), however the full signaling mechanism was not fully clarified. We therefore used GOAna to identify possible central modulators for the PGE_2-mediated inhibitory effect on activated T cells. In short, $CD4^+$ T cells were stimulated with CD3/CD28/MHC-I beads with or without PGE_2 for 4 days (**Figure 7.3A**). The Inhibitory effect of PGE_2 was demonstrated by a proliferation assay (**Figure 7.3B**), IFN-γ and TNF-α secretion (**Figure 7.3C**).

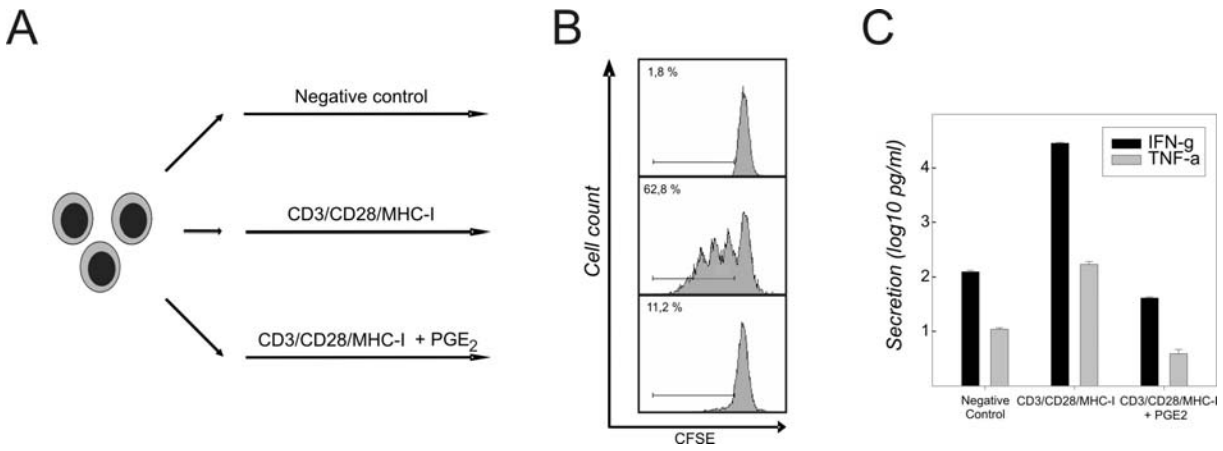


Figure 7.3 – Experimental setup
$CD4^+$ T cells were stimulated with CD3/CD28/MHC-I beads with or without PGE_2 for 4 days. CFSE dilution was determined using Flow Cytometry. Percentage of proliferating cells is shown. The amount of IFN-γ (black bars) and TNF-α (grey bars) secreted by the resting (negative control), activated (CD3/CD28/MHC-I) and PGE_2-treated (CD3/CD28/MHC-I + PGE_2) $CD4^+$ T cells was measured using a cytometric bead array. Shown are triplicates of one representative experiment.

When comparing activated $CD4^+$ T cells and activated $CD4^+$ T cell treated with PGE_2 using the GOAna algorithm, we identified 79 significant gene spaces (p-value < 0.05), including biological processes

concerned with 'RNA processing', 'apoptosis' and 'regulation of cell growth'. These processes clearly resemble the known biological differences between activated $CD4^+$ T cells and activated $CD4^+$ T cells in the presence of the inhibitory molecule PGE_2.

By iteratively increasing x (Equation (2)) we constructed several contrib-networks (**Figure 7.4**) and revealed that three subunits of protein phosphatase type 2A (PP2A) – PP2R1A, PPP2R1B and PP2CA – appeared in 24 of the 79 gene spaces (**Figure 7.4D**).

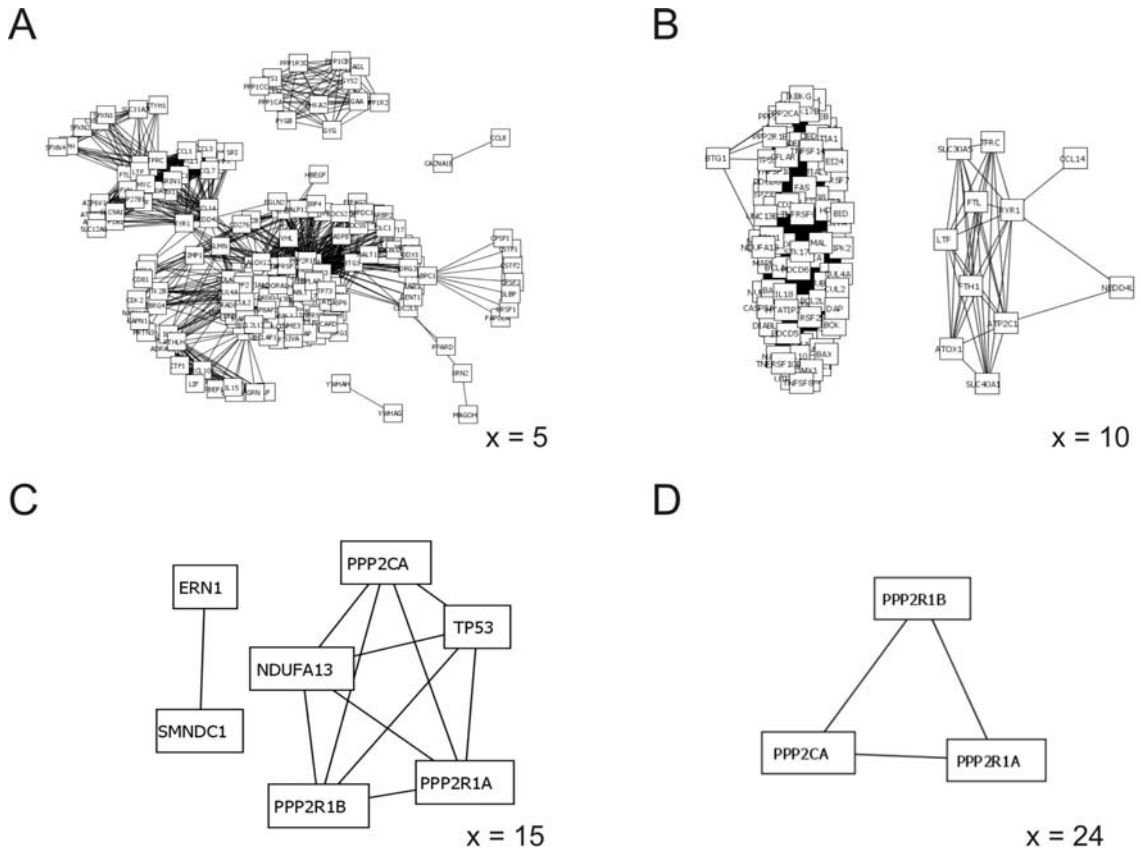


Figure 7.4 – Contrib-networks resulting from GOAna
Several contrib-networks were constructed by iteratively increasing x (Equation (2)). Shown are contrib-networks based on (**A**) x=5, (**B**) x = 10, (**C**) x = 15 and (**D**) x = 24.

PP2A has been identified as an important regulator of signal transduction and cell growth and functions by dephosphorylation of downstream targets, for example ERK, PKA and PKB and AKT (Garcia, Cayla et al. 2003; Van Hoof and Goris 2004; Mumby 2007). Indeed, no transcriptional regulation of PP2A was observed (**Figure 7.5A**). If PP2A in fact acts as an upstream central modulator in PGE_2 signaling, the phosphorylation status of downstream targets should alter based on the

presence of PGE_2. To interrogate this hypothesis, we performed a Western blot analysis using a Jurkat cell line and the phosphorylation status of ERK as readout.

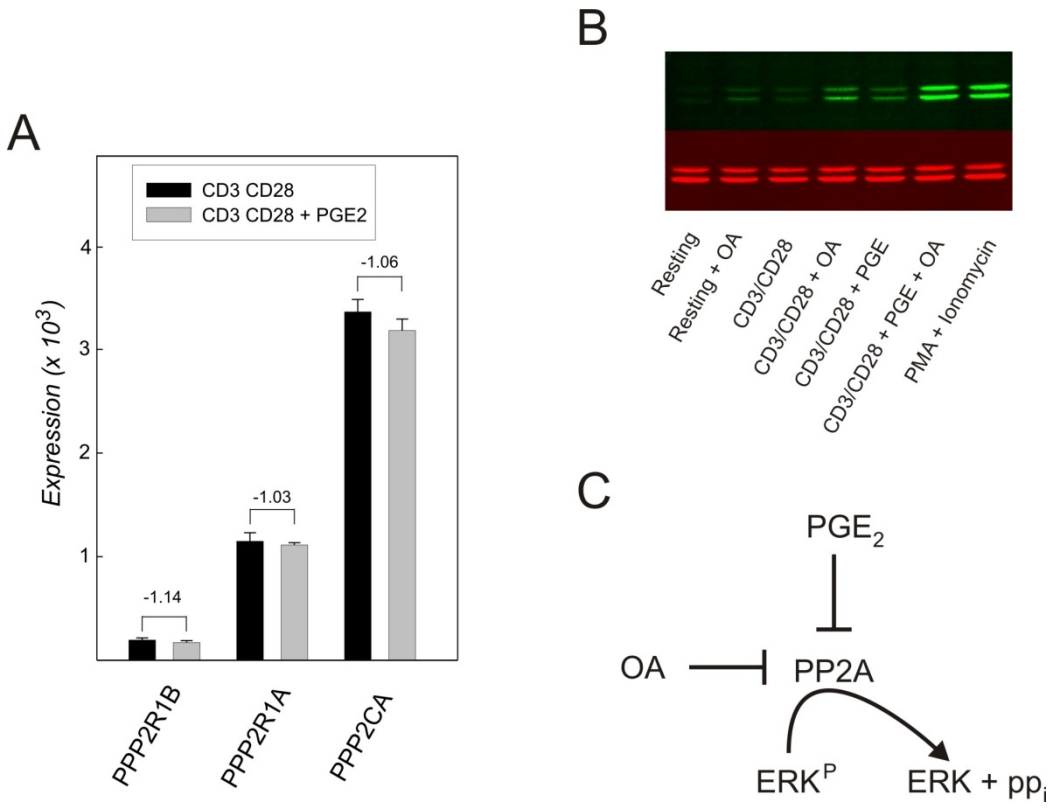


Figure 7.5 – Functional investigation of PP2A and proposed mechanism
(**A**) Expression of the three key players PPP2CA, PP2R1A and PPP2R1B identified by GOAna. Shown are expression values and fold changes between $CD4^+$ T cells stimulated with CD3/CD28/MHC-I (black bars) and $CD4^+$ T cells stimulated with CD3/CD28/MHC-I after addition of PGE_2 (grey bars). (**B**) Western Blot analysis using the phosphorylation status of ERK as readout. $CD4^+$ T cells were either left unstimulated (Resting) or were stimulated with CD3/CD28/MHC-I beads with (CD3/CD28 + PGE_2) or without PGE_2 (CD3/CD28). Ocadaic acid (OA), a known inhibitor of PP2A, was used as a positive control. As a further control, PMA + Ionomycin which directly phosphorylates ERK was used. Shown is the phosphorylated ERK (green bands) as well as total ERK (red bands).

Jurkat cell lines were stimulated with CD3/CD28/MHC-I beads in the presence or absence of PGE_2. Ocadaic acid, a known inhibitor for PP2A was used as a control. As a further control, PMA + Ionomycin which directly phosphorylates ERK was used.

As seen in **Figure 7.5B**, in the presence of PGE_2, phosphorylation of ERK was detected, although to a lesser extend as when treated with ocadaic acid (OA), a known inhibitor of PP2A. When treating activated Jurkat cell with both PGE_2 and OA, the phosphorylation of ERK was enhanced compared to treatment with PGE_2 or OA alone. Based on these results we propose PGE_2 to act as a repressor on PP2A (**Figure 7.5C**).

7.3 Discussion and further research options

Genome-wide transcriptional approaches have been extensively used to unravel the mode of action of diverse signaling processes. However, getting insight into signaling processes which are not reflected by transcriptional changes is still a challenging task. Here we propose, for the first time, a method which uses a whole genome-wide gene expression dataset as a RNA fingerprint to predict upstream events reflected by the observed transcriptional changes. Using this algorithm we determined PP2A as a key player which links several biological processes involved in the separation of activated T cells and activated T cells in the presence of PGE_2. Additionally, we confirmed this finding by Western blot analysis using a Jurkat cell line. Here PP2A was identified as central modulator in PGE_2 signaling which is repressed in the presence of PGE_2. Common approaches, as for example searching for differentially expressed genes or even pathways, would have not resulted in the identification of PP2A, since PP2A signaling depends on phosphorylation of target molecules. As the confirmation was performed in a Jurkat cell line and there are known difficulties in the comparability of Jurkat cell lines and primary T cells (Abraham and Weiss 2004), the logical consequence is performing the same experiment in primary $CD4^+$ T cells. Right now the Western blot analysis using ERK as readout is performed on freshly isolated $CD4^+$ T cells.

Part III: Further developments and critical considerations

8 IlluminaGUI - an application for establishing RNA fingerprints

Establishing and applying RNA fingerprints requires sophisticated data analysis methods for gene expression data. In this chapter I would like to introduce a software package which allows researchers to perform the above mentioned investigations using Illumina's Sentrix BeadChip technology (see Chapter 2.4.2).

8.1 Motivation

One of the most recent technologies in the area of genome-wide transcriptional profiling is the Sentrix BeadChip technology developed by Illumina (CA, USA) (Kuhn, Baker et al. 2004). Although the technology has been proven to be of highest quality (Patterson, Lobenhofer et al. 2006; Shi, Reid et al. 2006) widespread use by the novice as well as experienced life scientist is hampered due to the lack of comprehensive analysis tools specifically developed for this technology platform. For users of the Illumina BeadChip technology the options for sophisticated data analysis are currently limited to Illumina's BeadStudio or the Bioconductor packages 'beadarray' (Dunning, Smith et al. 2007), 'lumi' and 'BeadExplorer'. The Bioconductor project (Gentleman, Carey et al. 2004) - primarily based on R (R Development Core Team 2007) - is one of the most widely used open source software platforms for computing microarray data. The three Bioconductor packages mentioned are 'command line'-based and are designed for scientists with sufficient programming skills. While the BeadExplorer's limited GUI interface is restricted to quality control methods and data normalization, Illumina's BeadStudio tool offers basic analysis tools in a GUI environment, however, lacks many methods necessary for comprehensive microarray analyses including high-level statistical analyses. Moreover, it does not make use of widely accepted algorithms e.g. implemented in R. To overcome the current limitations of data analysis using the Illumina platform, I developed IlluminaGUI, a R package implementing a graphical user interface (based on the R-Tcl/Tk interface (Dalgaard 2001) for microarray data analysis. IlluminaGUI is designed specifically for life scientists who are not familiar with a command line based environment like the R language but do not want to resign the vast analysis opportunities of R. IlluminaGUI is freely available under http://IlluminaGUI.dnsalias.org.

8.2 Results

IlluminaGUI offers a collection of R functions combined to an easy-to-use GUI-based analysis tool covering the key components of a microarray analysis - preprocessing, inference and classification. As input files IlluminaGUI is using the primary data output files derived from Illumina's Beadstudio. These files are in tab-delimited format and include the non-normalized expression values together with detection p-values for each probe. To date, all available BeadChip versions (Human-Ref8, Human WG6v1, Human WG6v2, Mouse-Ref8, Mouse6v1, Mouse6v1.1, Rat-Ref12) are supported. Preprocessing of the data includes visualization of the data using basic diagnostic plots (e.g. MA plots (Dudoit, Yang et al. 2002), box plots (**Figure 8.1A**) and pairwise scatter plots), determination of absent/present genes as well as normalization of the data using state-of-the-art normalization techniques. Here, the quantiles-method (Bolstad, Irizarry et al. 2003), the vsn-method (Huber, von Heydebreck et al. 2002) and the qspline-method (Workman, Jensen et al. 2002) are implemented.

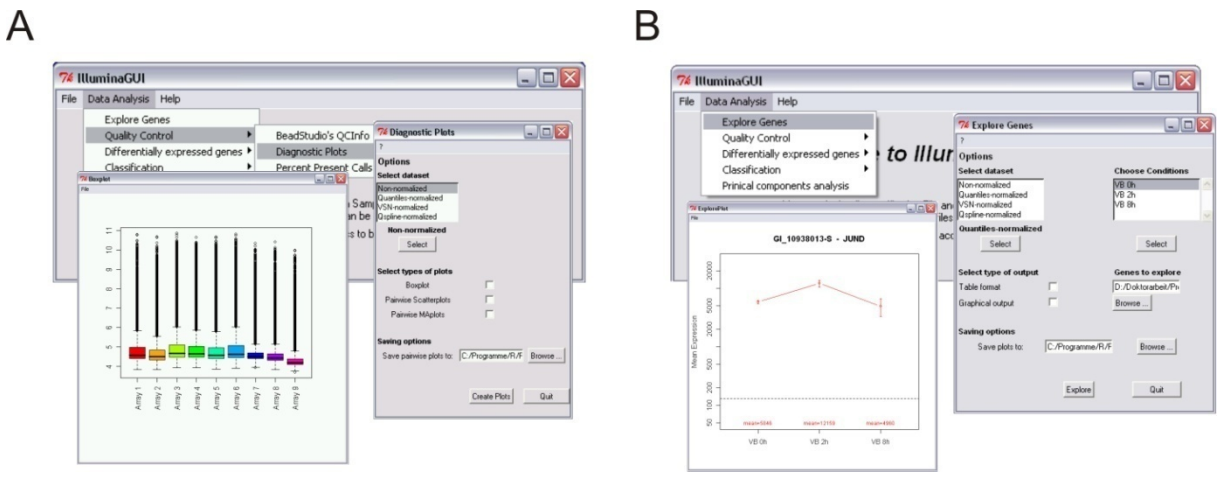


Figure 8.1 – IlluminaGUI visualization methods
(A) For quality control IlluminaGUI provides visualization of the data using diagnostic graphical plots. These include pairwise scatter plots, box plots and MA plots. Shown here is an example of a box plot of the data before normalization. **(B)** Explore plot.

We also introduce a new graphical tool - 'ExploreGenes' - which has, to our knowledge, not been described or implemented in any of the R packages yet. With 'ExploreGenes' the user can examine the expression profile for predefined genes across the entire dataset or parts of the dataset. The profile is displayed as a contour-plot showing mean values of biological replicates and can be exported to an Excel-file or as graphical output (**Figure 8.1B**).

IlluminaGUI provides several methods for identifying differentially expressed genes. Besides the combined t-test/fold change analysis, linear model analysis using LIMMA (Smyth 2004) is offered

to investigate the dataset. The methods are combined with procedures to correct for multiple testing, e.g. the FDR. In addition the user is able to perform SAM analysis (**Figure 8.2**) as described by Tusher et al. (Tusher, Tibshirani et al. 2001).

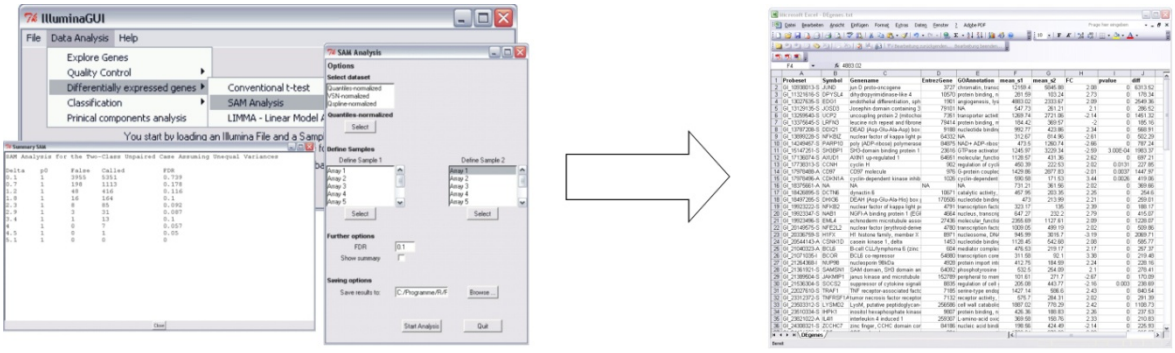


Figure 8.2 – IlluminaGUI inference methods
Analysis of differentially expressed genes is provided by three different methods, conventional t-test/fold change analysis, Linear Model analysis using LIMMA and SAM analysis. Shown here is the result of a SAM analysis. The result is also saved to an EXCEL-file.

All inference methods provide a fully annotated output file which includes probeset IDs, symbols, gene names, Entrez Gene IDs and a Gene Ontology annotation (**Figure 8.2**). For this purpose annotation packages based on the original annotation provided by Illumina have been created and will be distributed along with the IlluminaGUI package.

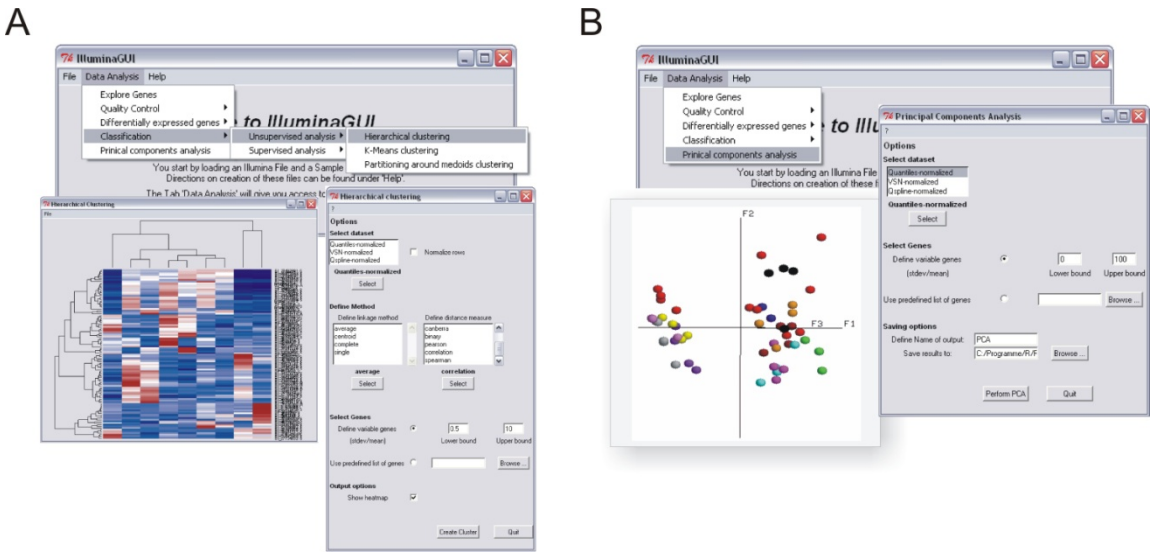


Figure 8.3 – IlluminaGUI classification methods and PCA
(A) Unsupervised classification provided in IlluminaGUI includes hierarchical clustering. Here, samples can be clustered based on a list of variable or pre-defined genes. **(B)** Principal components analysis in IlluminaGUI provides an html-output of the data which can be viewed using a common browser and the Chime plug-in (Chime Plugin 2008).

For classification purposes both unsupervised and supervised classification methods are offered. At present, unsupervised classification is provided in terms of hierarchical clustering, k-means clustering and partitioning around medoids clustering. For hierarchical clustering, various distance measures and linkage methods can be selected (**Figure 8.3A**).

Genes for clustering can be defined by filtering of variable genes based on normalized variance across samples or predefined gene lists. Supervised classification is performed using two different methods - nearest shrunken centroids (PAM) and support vector machines (SVMs). Here, different feature selection techniques and cross-validation are provided. With both, the supervised and unsupervised methods the user obtains written as well as graphical outputs of the results. For example, when using PAM, the accuracy of each prediction, the overall accuracy and the certainty of a prediction can be exported to an Excel-File. In addition, a probability plot displaying the result of all predictions is created. Similarly to hierarchical clustering, principal components analysis (PCA) of samples can be performed using variable genes or a predefined list of genes. Visualization of the results is provided as an html-output using a common browser and the Chime plug-in (Chime Plugin 2008). IlluminaGUI not only enables the user to save graphical outputs and analysis results, but also to save an entire project, i.e. the analysis can be continued at any time from the point of saving. This feature avoids tedious computations to be repeated all over when restarting or continuing an analysis.

8.3 Discussion and further research options

IlluminaGUI is a microarray analysis tool intended to enable the interested life scientist analyzing microarray experiments based on the increasingly used Illumina technology. In addition, IlluminaGUI can support the experienced user to expedite gene expression data analysis e.g. in a service lab environment. IlluminaGUI covers all aspects of a microarray experiment, starting from graphical quality controls to high-level statistical analyses as, for example, PCA or supervised classification. While IlluminaGUI will enable the life scientist to perform a basic microarray workflow without the help of experts in bioinformatics, at the same time, it is intended to enable the novice microarray user to achieve a rather sophisticated gene expression analysis as a basis for fruitful interactions with experts in statistics and bioinformatics. It is planned to extend IlluminaGUI in several different ways. Gene class testing approaches as introduced in Chapter 6.1 are more and important in analyzing microarray experiments. Different gene class testing approaches, including the Bioconductor package GOstats and the GOAna algorithm will be included in the interface. Additionally we are planning on extending the interface for analysis of high throughput miRNA data.

9 Critical considerations about underlying technology

In the following chapter I would like to focus on the challenges the underlying technology might implicate on the creation of RNA fingerprints. All described approaches for the creation of RNA fingerprints are heavily dependent on the reliability of the microarray format used for the study. Reasonable concerns would include the reproducibility of RNA fingerprints on different platforms or the continuity of RNA fingerprints when a new version of a microarray with updated probe content becomes available. The following study interrogates the latter concern and uses the Illumina BeadChip technology as an example to assess the impact of probe changes on the achieved results.

9.1 Motivation

Several reports have raised concerns about the comparability of microarray results coming from different platforms (Irizarry, Warren et al. 2005; Larkin, Frank et al. 2005; Kuo, Liu et al. 2006). However, recently the MicroArray Quality Control (MAQC) project has made a significant contribution assuring reliability and consistency of DNA microarray technology (Canales, Luo et al. 2006; Guo, Lobenhofer et al. 2006; Patterson, Lobenhofer et al. 2006; Shi, Reid et al. 2006; Shippy, Fulmer-Smentek et al. 2006; Tong, Lucas et al. 2006). The major message from the MAQC project, a community-wide effort initiated and led by FDA (US Food and Drug Administration) scientists, is that microarrays with comparable content show inter- and intra- platform reproducibility of gene expression measurements. Major regulatory agencies such as the FDA or the European Medicine Agencies (EMEA) have recognized genomic technologies, particularly gene expression profiling by DNA microarrays, as opportunities in advancing personalized medicine (Lesko and Woodcock 2004; Frueh 2006). Therefore, the results established by MAQC are very promising for the use of DNA microarrays in drug development, medical diagnostics and risk assessment, and the use of these technologies has been encouraged by the regulatory agencies. However, as already outlined by the MAQC project, an important aspect of DNA microarray technology needs further attention (Shi, Reid et al. 2006). Advances in array technology as well as improvements of genomic database content will lead to the development of new generations of microarrays in upcoming years (Hardiman 2006; Hoheisel 2006). The currently available annotation of transcripts represented on DNA microarrays (microarray content) is still incomplete. In fact, our knowledge about gene expression is far from being complete, which is reflected by a continuous increase of content of gene databases such as RefSeq (Pruitt, Tatusova et al. 2007). Moreover, still more than 50% of Human RefSeq entries are

only preliminarily annotated (Johnson, Castle et al. 2003). Starting with a few hundred transcripts a decade ago current versions of DNA microarrays interrogate transcripts in the order of 50,000. So far, using the most recent DNA microarray technology has always been seen as an advantage - especially when searching for novel transcripts (Classen, Zander et al. 2007). However, this might be different in the setting of drug development, medical diagnostics or risk assessment, where patterns of expression rather than single genes are of highest relevance. Here, permanent gene annotation and probe sequence content are needed for long-term applications. The potential impact of advances in technology and database content on successfully established diagnostic gene signatures (e.g. the 70-gene signature established by van't Veer et al. for predicting therapy outcome in breast cancer patients (van 't Veer, Dai et al. 2002; van de Vijver, He et al. 2002) or the RNA fingerprint predicting a lung cancer incidence (see 1)) has not been fully appreciated. It is therefore mandatory to develop approaches and methods that allow fast and decisive assessment of the global impact database improvements, content changes of microarrays and technical advances might impose on the use of DNA microarray technology.

9.2 Dealing with next generation microarrays: A solution strategy

As a consequence, fast, reliable and standardized assessment of technological advances in array technology and content is critically needed leading us to propose a methodology that allows assessment of

1. The amount of changes on a probe content level between subsequent versions of microarrays
2. The technological improvements between subsequent versions and
3. The impact of these improvements on reproducibility and comparability of biological results.

We therefore describe a methodology allowing rapid determination of the impact of introducing newer generation microarray technology with improved genomic content on gene expression analysis results. This method consists of *in-silico* analyses of microarray content combined with a performance analysis using real biological samples.

Significant dynamics of gene sequence content of current genome databases

One of the major resources for genomic research are databases such as RefSeq (Pruitt, Tatusova et al. 2007), Unigene (Pontius, Wagner et al. 2003), Ensembl (Flicek, Aken et al. 2007), or GenBank

(Benson, Karsch-Mizrachi et al. 2006). Due to the enormous gene cloning efforts during the last years, the content of gene databases is dramatically increasing. Plotting the official release statistics of the RefSeq database (ftp://ftp.ncbi.nih.gov/refseq/release/release-statistics/) shows the continuing growth of gene RefSeq sequences (**Figure 9.1A**) mainly explained by constant addition of new species.

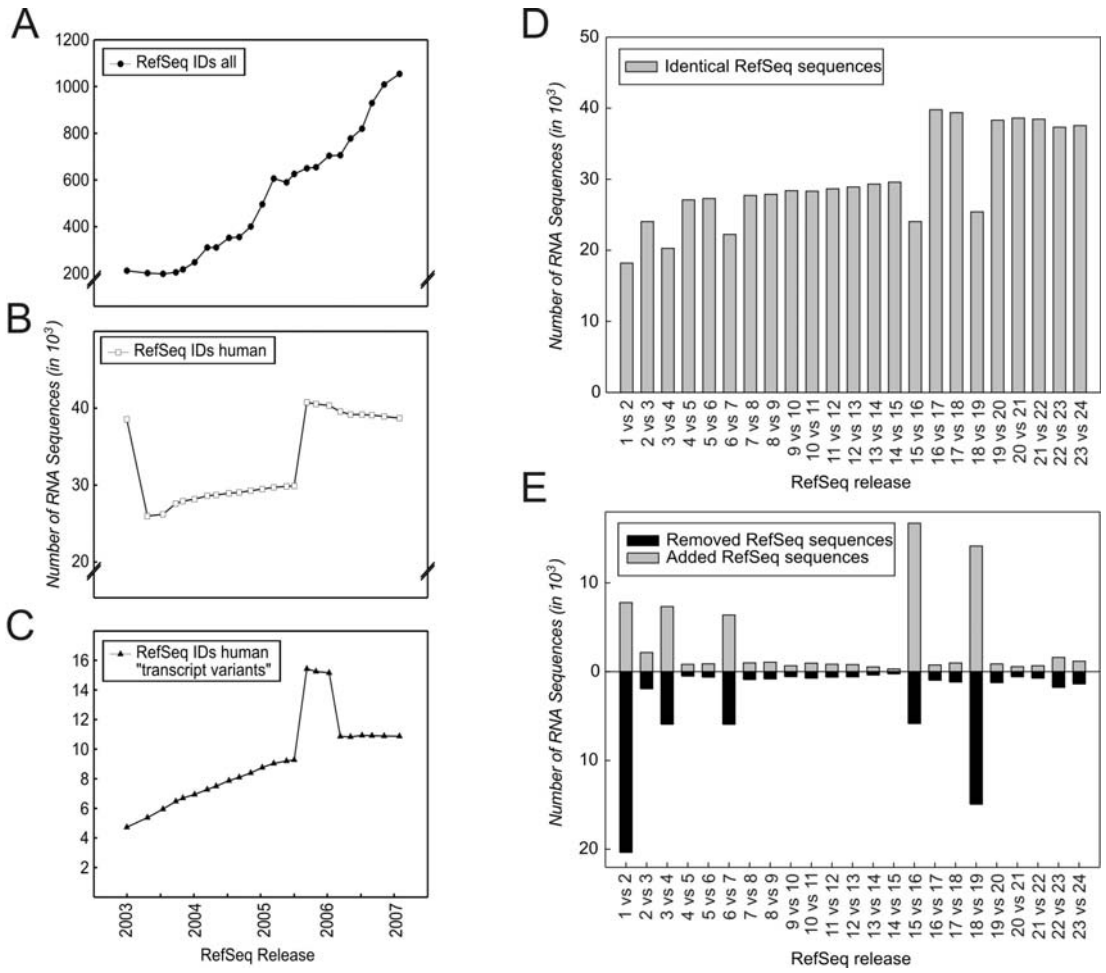


Figure 9.1 – Dynamics of RefSeq database
Official release statistics retrieved from (ftp://ftp.ncbi.nih.gov/refseq/release/release-statistics/) shows the development of the RefSeq database, including (**A**) all RefSeq IDs, (**B**) human RefSeq IDs, and (**C**) human RefSeq IDs termed "transcript variant". Consecutive releases were compared to each other to obtain (**D**) concordances and (**E**) changes in the database over time.

One of the latest versions of RefSeq (September 14, 2007) covers 6,515,158 entries coding for 4,167,224 proteins from a total of 4,646 organisms. To determine the development of the content of human gene sequences (human database entries, huDE) huDE from the RefSeq release catalog were extracted (ftp://ftp.ncbi.nih.gov/refseq/release/release-catalog). Starting with almost 40,000 huDE in release one (R1) the content dropped to less than 28,000 huDE, steadily increased to 30,000 huDE

(R16) when almost 11,000 huDE were added in R17. Since then the overall number of huDE remained stable (**Figure 9.1B**). The increase of huDE observed from R2 to R17 is mainly explained by new knowledge concerning transcript variants (mainly splice variants). Transcript variants have been added continuously to the database (**Figure 9.1C**). Since 2003 the number of known splice variants more than doubled reaching now 10,000 huDE (R24). While reaching a plateau in overall content of huDE (**Figure 9.1D**), assessment of absolute numbers does not necessarily reflect additional dynamics of the database due to exchange of huDE. We therefore assessed changes of huDE between subsequent releases over time. This analysis revealed a surprisingly high number of changes between subsequent releases, even for the latest releases (**Figure 9.1E**). These changes can be explained by constant curation of the database including nucleotide changes of existing sequences, removal of redundant or non-informative content and addition of newly identified sequences like splice variants. Based on these unexpected and still high dynamics of human genome content, we hypothesized that the broadly applied microarray technologies, for which RefSeq is one of the main repositories, will be strongly influenced by such changes.

Content and annotation of microarrays depends on the reference database

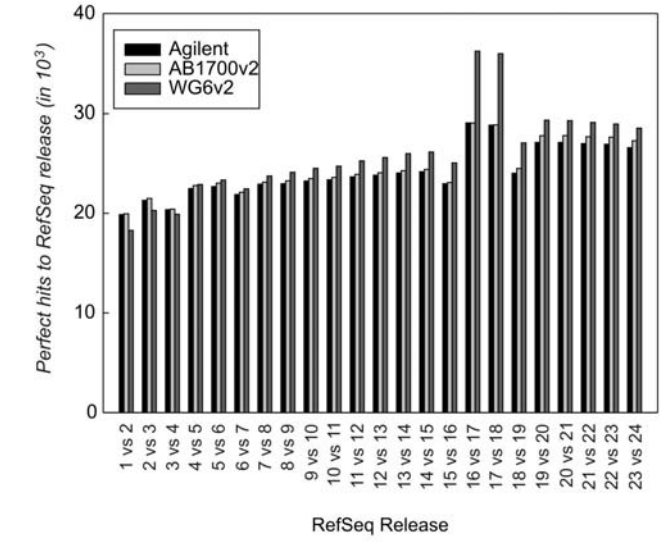
To address the influence of database content on array design and layout we first assessed the impact of different RefSeq releases on array annotation. As a model we used three commercially available oligonucleotide-based microarray platforms, the Whole Human Genome Oligo Microarray distributed by Agilent (A-huGOM), the Human Genome Survey Microarray distributed by Applied Biosystems (AB-huGSM) and the Human BeadChip distributed by Illumina (I-huBC). All three microarray systems are based on long oligonucleotides (≥ 50-mers) which are used for hybridization to their specific transcripts (**Figure 9.2A**) and are known to have RefSeq as one of their major sequence repository (Applied Biosystems; Kronick 2004; Kuhn, Baker et al. 2004).

For this analysis the most recent versions of the respective microarrays were used. All oligonucleotide probes present on the microarray were blasted against RefSeq releases R1 to R24 to determine the proportion of annotated probes on the respective array. As can be seen in **Figure 9.2B**, all three microarray formats show similar constant levels of annotation for subsequent releases. The large change in sequence content observed in RefSeq database R16 (see **Figure 9.1B**) also showed a huge increase of the number of annotated probes on the microarrays with the most prominent rise observed for the I-huBC. Again, when investigating annotation changes between subsequent releases for all three microarray platforms (**Figure 9.2C**), the pattern of annotation

changes showed a high similarity to the pattern of database changes (**Figure 9.1E**) which reflects a high correlation of database content and annotated probes on microarrays.

A

Array name	Distributor	Features	Feature type
Whole Human Genome Oligo Microarray	Agilent	32787	60-mer
Human Genome Survey Microarray	Applied Biosystems	41000	60-mer
Human-6 Expression BeadChip	Illumina Inc.	48701	50-mer



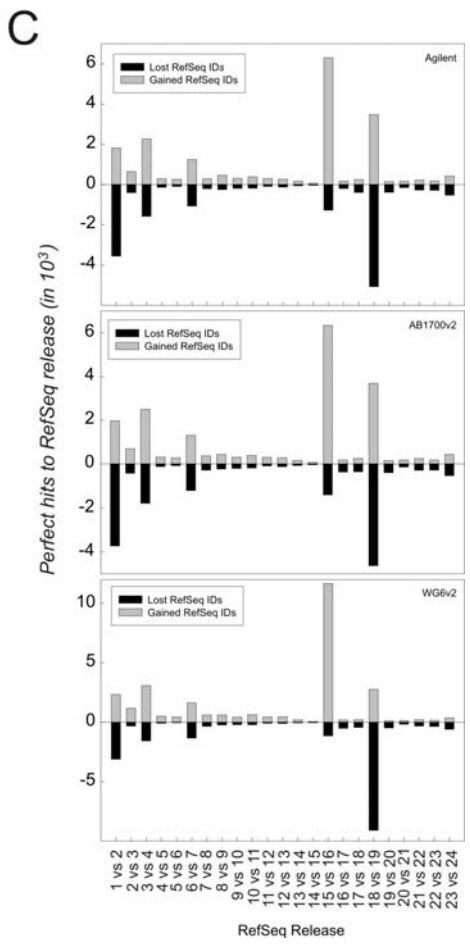


Figure 9.2 – Influence of Refseq database content on annotation of microarray probes
(**A**) Array type, feature type and number of features interrogated by three commercially available oligonucleotide-based microarray platforms. (**B**) Influence of RefSeq version on annotation of probes used by the three microarray platforms. (**C**) Differences in the annotation status based on differences of consecutive Refseq versions for the A-huGOM, the AB-huGSM and the I-huBC.

The high dynamics in database content and subsequent annotation changes result in the need for constant update of probe content on microarrays. We therefore were particularly interested in characterizing the impact of content by comparing different generations of microarrays developed on the basis of different database content.

Consistency of consecutive array versions strictly depends on database content and annotation

To interrogate the impact of database content changes we investigated both the Human BeadChip distributed by Illumina (I-huBC) and the Human Genome Survey Microarray distributed by Applied Biosystems (AB-huGSM). Both companies recently launched a second version of their original product: AB-huGSM-V2, released in January 2005 and I-huBC-V2, released in December 2006. These enabled us to examine changes in probe content between subsequent array releases. The two arrays distributed by Applied Biosystems are comprised of 33,096 (AB-huGSM-V1) and 32,787 (AB-huGSM-V2) oligonucleotides, respectively. 30,469 oligonucleotides were identical between version 1 and 2, whereas 2,627 oligonucleotides were removed and 2,318 oligonucleotides were added to adapt the new array format to changes in the RefSeq database (**Figure 9.3A**).

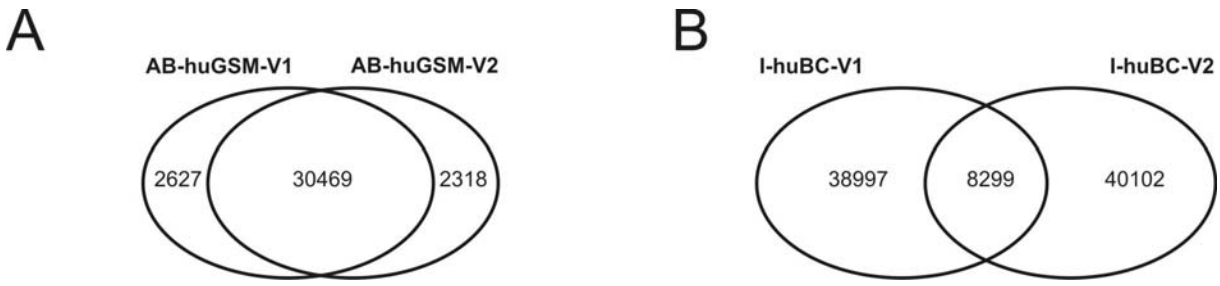


Figure 9.3 – Comparison of probe level content on subsequent array versions
For (**A**) the AB-huGSM and (**B**) the I-huBC two subsequent array versions were compared regarding their probe level content.

The Illumina BeadChip arrays included 47,296 (I-huBC-V1) and 48,401 (I-huBC-V2) probes, respectively (Kuhn, Baker et al. 2004), but to our surprise, only 8,299 oligonucleotides were identical between I-huBC-V1 and I-huBC-V2 (**Figure 9.3B**). This massive change in probe content from one array version to the next led us to more closely examine the differences in probe content of the I-huBC arrays. We postulated that comparability of array results is greatly challenged by introducing significant changes in probe content. To address this issue in detail, we assessed the overall magnitude of changes using I-huBC-V1 (version 1) and I-huBC-V2 (version 2) as a model. Generally, probe sequence changes on consecutive array versions can lead to different numbers and types of RefSeq hits in both array versions. Generally, probe sequence changes on consecutive array versions can lead to different numbers and types of RefSeq hits in both array versions. Types of RefSeq hits include “perfect” (100% sequence identity), “imperfect” (>90% sequence identity) or “unspecific” (<90% sequence identity) hits. Furthermore, number and type of RefSeq hits depend on changes within the RefSeq database.

We categorized RefSeq hits resulting from probe sequence changes as follows:

- Hit Category 1: RefSeq hit obtained by identical probe sequences represented on both array versions
- Hit Category 2: RefSeq hit obtained by distinct probe sequences (sequence changes in RefSeq, design improvement, etc.)
    - Category 2a: Hit to the same RefSeq ID(s) by distinct probe sequences
    - Category 2b: Perfect RefSeq hit on version 1, imperfect RefSeq hit on version 2
    - Category 2c: Perfect RefSeq hit on version 1, unspecific RefSeq hit on version 2
    - Category 2d: Imperfect RefSeq hit on version 1, perfect RefSeq hit on version 2
    - Category 2e: Unspecific RefSeq hit on version 1, perfect RefSeq hit on version 2
- Hit Category 3: New RefSeq is added (splice variants, prediction (XM_ probe) was correct)
- Hit Category 4: RefSeq is deleted (prediction (XM_ probe) turned out to be wrong, problems in synthesis, not important!)

We used this categorization to assess the impact of probe sequence changes on the comparability of the consecutive array versions. To interrogate differences in RefSeq hit categories between I-huBC-V1 and I-huBC-V2 we initially performed a BLAST analysis on all oligonucleotide sequences from both arrays using three RefSeq releases. R24 represents the actual release, R17 the release at the time of I-huBC-V2 array design, and R4 the release at the time of I-huBC-V1 array design.

In short, oligonucleotides from both array versions were blasted against the respective RefSeq release and hits which were called perfect were grouped into one of the 4 described categories. The obtained perfect hits by each array version and the distribution of hits to the respective categories are displayed for RefSeq Versions R4 (**Figure 9.4A**), R17 (**Figure 9.4B**), and R24 (**Figure 9.4C**). The BLAST analysis performed on R17 (**Figure 9.4B**) obtained the highest number of perfect hits for I-huBC-V2 (36,405) as well as the highest number of shared RefSeq hits between I-huBC-V1 and I-huBC-V2 (27,090). Also for this release the lowest number of removals (categories 4, 2b and 2c) as well as the highest number of additions (categories 3, 2d and 2e) was obtained.

Surprisingly, these numbers changed dramatically when performing the BLAST analysis on the most recent release R24 (**Figure 9.4C**). Both the number of obtained perfect hits for I-huBC-V2 and the number for shared RefSeq hits dropped. Also the number of removals increased and the number of additions decreased. This result clearly reflects the strong dependence of array content on database content used for array design. When performing the comparison between I-huBC-V1 and I-huBC-V2 based on R4 (**Figure 9.4A**) we observed the least agreement in probe level content, as well as the lowest gain of content and the highest number of removals.

The observed differences concerning the concordance of probe level content between I-huBC-V1 and I-huBC-V2 based on three different RefSeq releases raised the question whether the concordance would achieve an optimum.

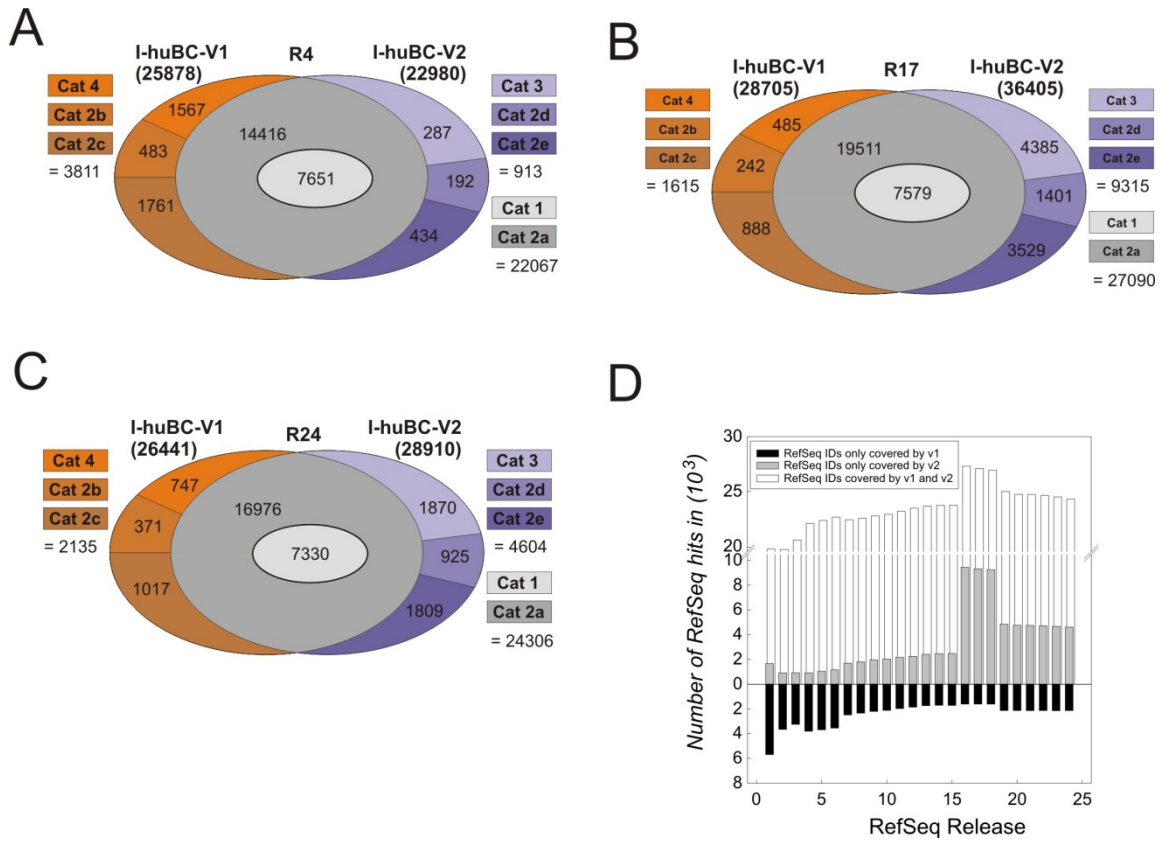


Figure 9.4 – Comparison of probe level content on subsequent array versions
The I-huBC-V1 and the I-huBC-V2 arrays were investigated regarding RefSeq hit categories in the following Refseq releases: (**A**) R4, (**B**) R17 **(C)**, and (**D**) R24. (**F**) Concordances and differences in probe level content between the I-huBC-V1 and the I-huBC-V2 array over all RefSeq releases.

Running the BLAST analysis on all official RefSeq releases (R1 to R24) revealed that concordance between I-huBC-V1 and I-huBC-V2 reached an optimum at R16 and R17 (**Figure 9.4D**), the existent releases at the time of array design of I-huBC-V2.

Number of cross-annotated probes on consecutive microarrays stays stable

For further analyses concerning performance issues of two different array versions we cross-annotated the re-blasted probes from I-huBC-V1 and the I-huBC-V2 arrays principally using the approach used by the MAQC project (Shi, Reid et al. 2006). In contrast to the MAQC project, which condensed its mapping to a 'one-probe-to-one-gene' approach, we took all perfect hits into account. Therefore, our cross-annotation approach had to consider three types of probes: (1) probes

which show a single perfect hit to a Refseq, (2) probes with multiple perfect hits to more than one Refseq which are all splice variants of the same gene and (3) probes which show hits to more than one Refseq comprising different genes. For each probe on the I-huBC-V1 array we compared its list of perfect Refseq hits to all probes on the I-huBC-V2 array. When identifying a probe on the I-huBC-V2 showing an exact match in length and content of the hit list, the two probes were cross-annotated. This approach ensured cross-annotation of probes within one probe type (1 to 3) but also excluded probes of type (2) which showed multiple hits on both I-huBC-V1 and the I-huBC-V2 but had distinct number of hits for both versions (distinct number of splice variants). In the latter case signals may not be comparable due to different expression profiles of single splice variants which would introduce further variation when investigating comparability of performance of two consecutive microarrays. Using this approach we cross-annotated probes based on all 24 RefSeq releases. We postulated that the number of cross-annotated probes would increase over time due to an increase of previously non-annotated probes hitting Refseq IDs in later versions of Refseq. To our great surprise, the number of cross-annotated probes stayed relatively constant over all releases (**Figure 9.5**) and did not show the expected increase at R16 and R17.

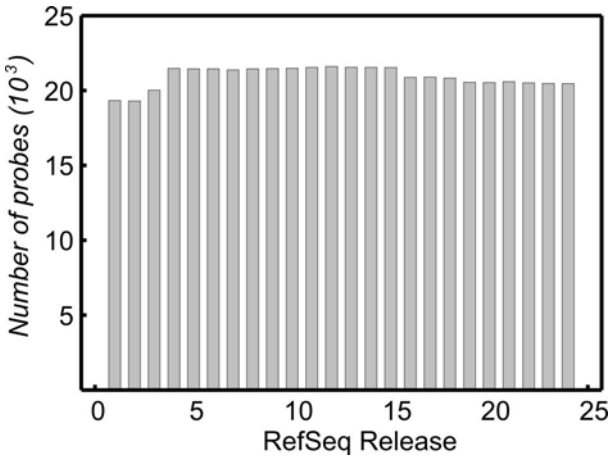


Figure 9.5 – Cross-annotation of probes
For each release probes from I-huBC-V1 were cross-annotated to probes from I-huBC-V2. For each probe I-huBC-V1 its list of perfect Refseq hits was compared to all probes on I-huBC-V2. Two probes were cross-annotated when identifying a probe on the I-huBC-V2 which showed an exact match in length and content of the hit list. Depicted is the number of cross-annotated probes for each release.

This can be explained by the huge increase in the number of splice variants for the two releases (see **Figure 9.1C**) as well as the cross- annotation approach itself which prohibits cross-annotation of probes targeting a distinct number of splice variants. For all further analyses we used the cross-annotation based on the latest RefSeq release (R24) and therefore worked on 20,456 probes.

Altogether, comparability of consecutive array versions even on a single platform is a function of oligonucleotide design, database content and annotation available at the time of array design. Unexpectedly, optimal comparability is not achieved with the newest annotation of the RefSeq database but rather with the annotation available at the time of design of the newest array version. As long as the database content is not yet finalized, updates in array design are mandatory to correctly reflect genomic content.

Selection of data sets for best investigation of performance issues

The above described in silico analysis of consecutive array designs (based on updated database releases) is an important first step to estimate the overall impact on array performance. However, we postulate that site-by-site comparison of performance of consecutive array versions by applying biological experiments is the most critical part of future array development as well as compatibility analysis for long-term projects spanning the life time availability of different array versions. Conceptually, these guiding experiments should fulfill the following criteria: representative data sets to assess array performance in (1) a biological screening experiment (e.g. cell type comparison) respectively (2) in a group analysis setting (e.g. clinical sub-classification of diseases), (3) coverage of as many present probes as possible, and (4) availability of validating data for single genes. To achieve these goals we performed two different sets of experiments. As an example for a biological screening experiment we compared conventional $CD25^+$ $CD127^-$ regulatory T cells (T_{reg}, n=3) as a specialized T cell subpopulation and compared these with so-called conventional $CD25^-$ $CD127^+$ T cells (T_{conv}, n=3) (**Figure 9.6**) (Liu, Putnam et al. 2006; Seddiki, Santner-Nanan et al. 2006).

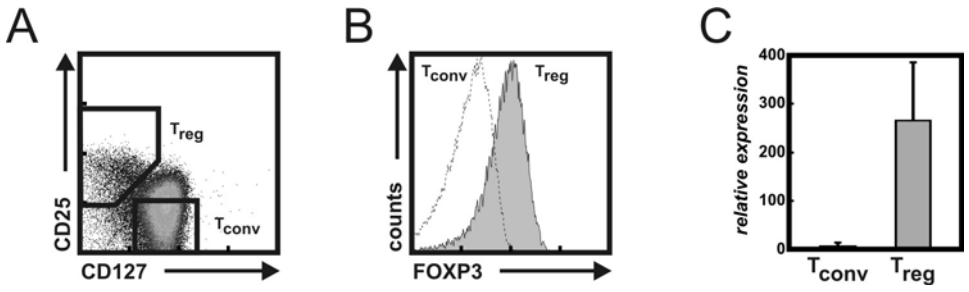


Figure 9.6 – Quality assessment of T_{reg} cells
Confirmation of the Treg cell population was performed using FACS analysis and sorting gates for $CD4^+$ CD127low $CD25^+$ T_{reg} cells and $CD4^+$ $CD127^+$ $CD25^-$ T_{conv} cells (**A**) as well as expression of FOXP3 in the respective T cell subsets assessed by flow cytometry (**B**) respectively quantitative RT- PCR (**C**).

$CD4^+$ T cells, isolated from peripheral blood were cell-sorted based on CD25 and CD127 expression into $CD25^+$ $CD127^-$ T_{reg} cells and $CD25^-$ $CD127^+$ T_{conv} cells (**Figure 9.6A**). Intracellular staining with

FOXP3 mAbs confirmed that $CD25^+$ $CD127^-$ cells were indeed T_{reg} cells (**Figure 9.6B**). Moreover, quantitative RT-PCR for the FOXP3 mRNA revealed high level expression of FOXP3 in $CD25^+$ $CD127^-$ T_{reg} cells but not $CD25^-$ $CD127^+$ T_{conv} cells (**Figure 9.6C**).

As a second set of experiments we chose a subgroup analysis of peripheral blood samples derived from patients with either scleroderma (n=9) or bacteremia (n=7). These samples are part of a larger study addressing diagnostic signatures of systemic diseases in peripheral blood (S. Debey-Pascher, unpublished results). Since transcriptional programs in peripheral blood differ significantly between scleroderma and bacteremia, we could restrict the comparison of the two consecutive arrays I-huBC-V1 and I-huBC-V2 to a smaller subset of samples. For all samples, we performed microarray analysis on both array types. Overall, the number of probes present in at least one sample (resp. sub-group) was 28,358 (resp.13,104), representing 72.9% (resp. 46.7%) of cross-annotated probes on the I-huBC-V1. The number of probes present in at least one sample (resp. subgroup) on I-huBC-V2 was 24,986 (resp. 18,096), representing 67.0% (56.0%) of cross-annotated probes, The larger number of probes called present in individual samples on the I-huBC-V1 array is also indicative for a higher variability of the earlier array version.

The new I-huBC-V2 outperforms the I-huBC-V1 array concerning sensitivity, signal-to noise-ratio and dynamic range

To quickly assess improvement of performance by newer generation technology, we assessed 4 parameters describing important quality aspects, namely (1) the percentage of detected transcripts reflecting sensitivity, (2) the dynamic range of signal intensities, (3) the values of background/noise signals reflecting signal-to-noise ratio and (4) technical replication reflecting reproducibility.

To investigate sensitivity we used the detection p-value to classify a probe as absent or present. In the T_{reg} data set on average 24.6% of all probes on I-huBC-V1 were called present, while on average 31.0% of all probes were called present on the I-huBC-V2 array. Similarly, in the peripheral blood data set, we obtained mean percentages of 23.2% for I-huBC-V1 and 30.1% for I-huBC-V2 samples. Additionally, probes with low signal intensities on both arrays were generally more often called present on I-huBC-V2 in comparison to I-huBC-V1 suggesting that the more recent array version has a lower detection limit. Next we determined the percentage of identical and cross-annotated probes called present on the I-huBC-V1 array, but absent on the I- huBC-V2 array and vice versa for each subgroup in the data sets (**Table 9.1**).

As expected from the higher percentage of probes called present on the I-huBC-V2 array, we also saw a significantly higher number of cross-annotated probes present on the I-huBC-V2 array

compared to the I-huBC-V1 array (up to 4 fold). Still, there was a small percentage of probes that were present on I-huBC-V1 but absent on I-huBC-V2.

		identical oligos		cross-annotated oligos	
		v1 P & v2 A	v1 A & v2 P	v1 P & v2 A	v1 A & v2 P
T_{reg} data set	T_{reg}	43 (0.5%)	544 (6.6%)	814 (4.0%)	2766 (13.5%)
	non T_{reg}	47 (0.6%)	445 (5.4%)	755 (3.7%)	3068 (15.0%)
Whole blood data set	Scleroderma	107 (1.3%)	250 (3.0%)	627 (3.1%)	3559 (17.4%)
	Bacteremia	82 (1.0%)	349 (4.2%)	762 (3.7%)	3207 (15.7%)

Table 9.1 – Absent respectively present status of probes
For each subgroup in the data sets the percentage of identical and cross-annotated probes called present on the I-huBC-V1 array, but absent on the I- huBC-V2 array (v1 P & v2 A) and vice versa (v1 A & v2 P) was determined.

Since we observed a rather high variability of probes called present in single samples compared to sub-groups on the I-huBC-V1, we hypothesized that these probes would have very low signal values (just above background value) on the I-huBC-V1 array and would therefore have been false-positively called present on the I-huBC-V1. Indeed, when determining these probes, ~75% showed values close to background level (data not shown). Altogether, these data further support that the newer generation array technology is of higher sensitivity respectively lower detection limit.
Boxplots can not only be used to determine the distribution of intensity signals across an array but also to compare the dynamic range of signals between different array types.

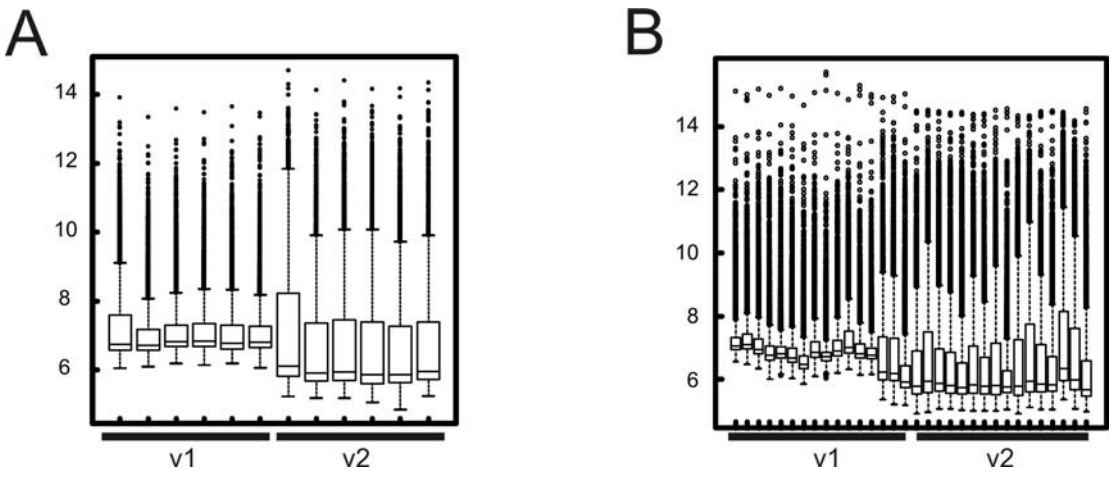


Figure 9.7 – Boxplots to determine the dynamic range of signal intensities
Boxplots were used to compare the dynamic range of signals on the arrays for (**A**) the Treg data set and (**B**) the whole blood data set. Only signals for the 8299 identical oligonucleotides were used.

When plotting the signals of the 8,299 probes that were identical on both array versions, we observed an enlargement of the dynamic range in I-huBC-V2 in both data sets (**Figure 9.7A, B**). Additionally a decrease in median signal intensities was observed which was due to reduced overall background values on the I-huBC-V2. This approach can easily be adapted to either compare all signals between two arrays or a subset of cross-annotated probes.

At least for identical oligonucleotide probes performance of a quidproquo technical replication between different array versions can be assessed on a sample-by-sample basis. When comparing raw signal intensities of such technical replicates we observed increased signal intensities for moderate to highly expressed transcripts on the I-huBC-V2 (**Supplementary Figure 1**). For visualization we used pairwise scatterplots, principal components analysis (PCA) and hierarchical clustering on normalized data.

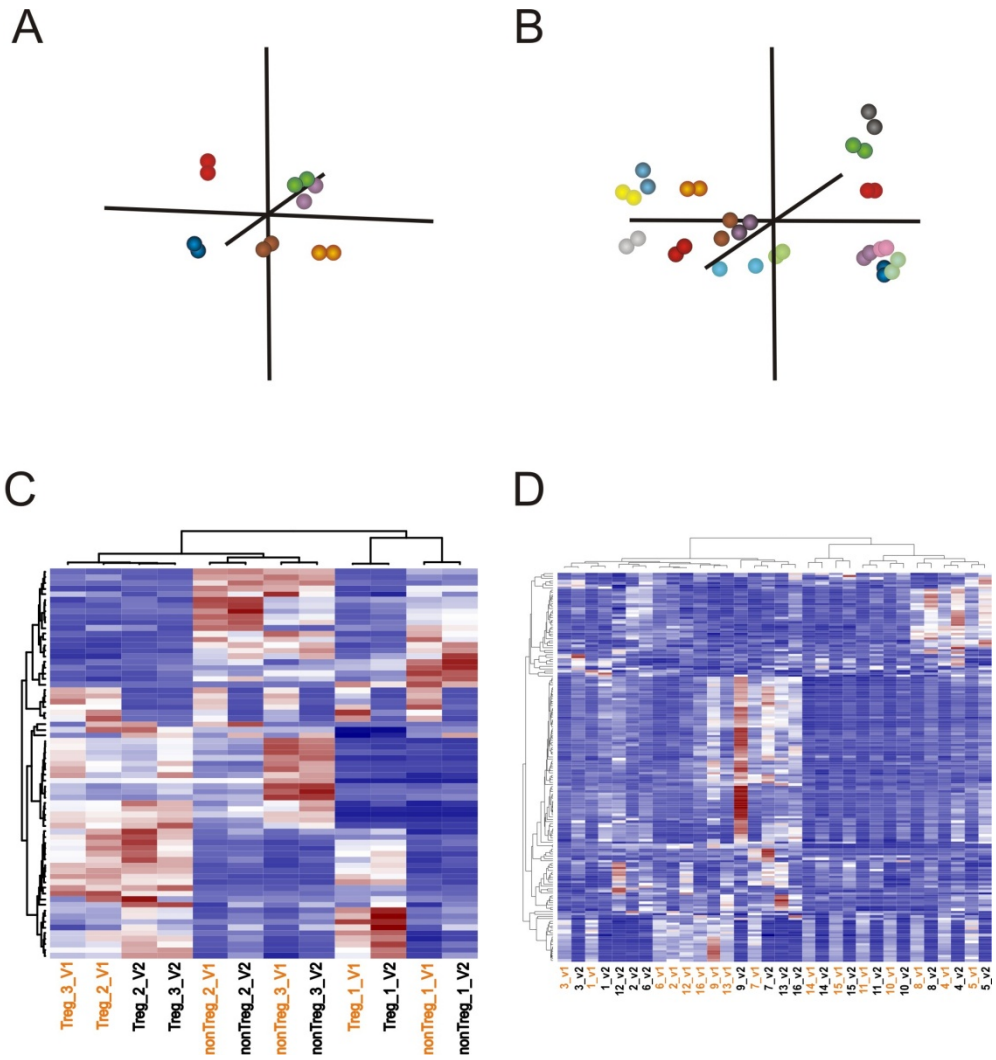


Figure 9.8 – Technical replication is assessed by PCA and hierarchical clustering
Technical replicates were checked using principle component analysis (PCA) based on the 100 most variable genes for (**A**) the Treg data set and (**B**) the whole blood data set and hierarchical cluster analysis of samples from (**C**) the Treg data set and (**D**) the whole blood data set.

For perfect technical replicates one would expect a straight diagonal line in pairwise scatterplots, side-by-side clustering of samples when applying PCA or clustering approaches, and a high pairwise correlation value. In fact, samples of the T_{reg} data set showed a mean correlation of 0.97 ± 0.005 and samples of the whole blood data set a mean correlation of 0.91 ± 0.17 (**Supplementary Tables 1 and 2**) which was visualized using pairwise scatterplots (**Supplementary Figure 2**). These results were confirmed when performing PCA using the 100 most variable probes out of the 8,299 identical oligonucleotides. When plotting the first three principal components of each sample in a 3D scatterplot a perfect side by side plot of technical replicates was observed (**Figure 9.8A**, **B**).

Additionally, we performed hierarchical clustering on these samples. Almost all technical replicates clearly clustered next to each other (**Figure 9.8C**, **D**). Altogether, the analysis revealed that technical replication using the more recent I-huBC-V2 array revealed comparable results concerning signal intensities.

Rank correlation metric reveals significant differences between subsequent microarray versions

Using a rank correlation metric is a common procedure to examine the comparability of results across platforms (Shi, Reid et al. 2006). We followed the approach taken by the MAQC project and used the ratio of differential expression (between defined groups, here T_{reg} versus T_{conv} resp. scleroderma versus bacteremia samples) as a basis for ranking transcripts between the I-huBC-V1 and the I-huBC-V2 array.

Ideally, highly comparable results would show a rank correlation close to 1. In a first step we used transcripts, which were moderately to highly expressed (signal intensity > 500) in either one of the sub-groups of the data sets to eliminate possible impairment due to absent or low expressed transcripts. **Figure 9.9A** shows the result of the analysis based on the 8,299 identical oligonucleotides in the T_{reg} data set. Here, 252 transcripts were moderately to highly expressed throughout the dataset and obtained a rank correlation of 0.95. When using the cross-annotated probes (628) the rank correlation dropped slightly to 0.85 (**Figure 9.9B**), which can most probably be ascribed to the differences in oligonucleotide placement in the gene (e.g. closer to 5`end). To our surprise, this high comparability could not be achieved for the whole blood data set. Using highly expressed identical oligonucleotides (99) we obtained a rank correlation of 0.77 (**Figure 9.9C**). In contrast to the T_{reg} data set this rank correlation remained constant (0.78) when performing the analysis on 269 cross-annotated (**Figure 9.9D**).

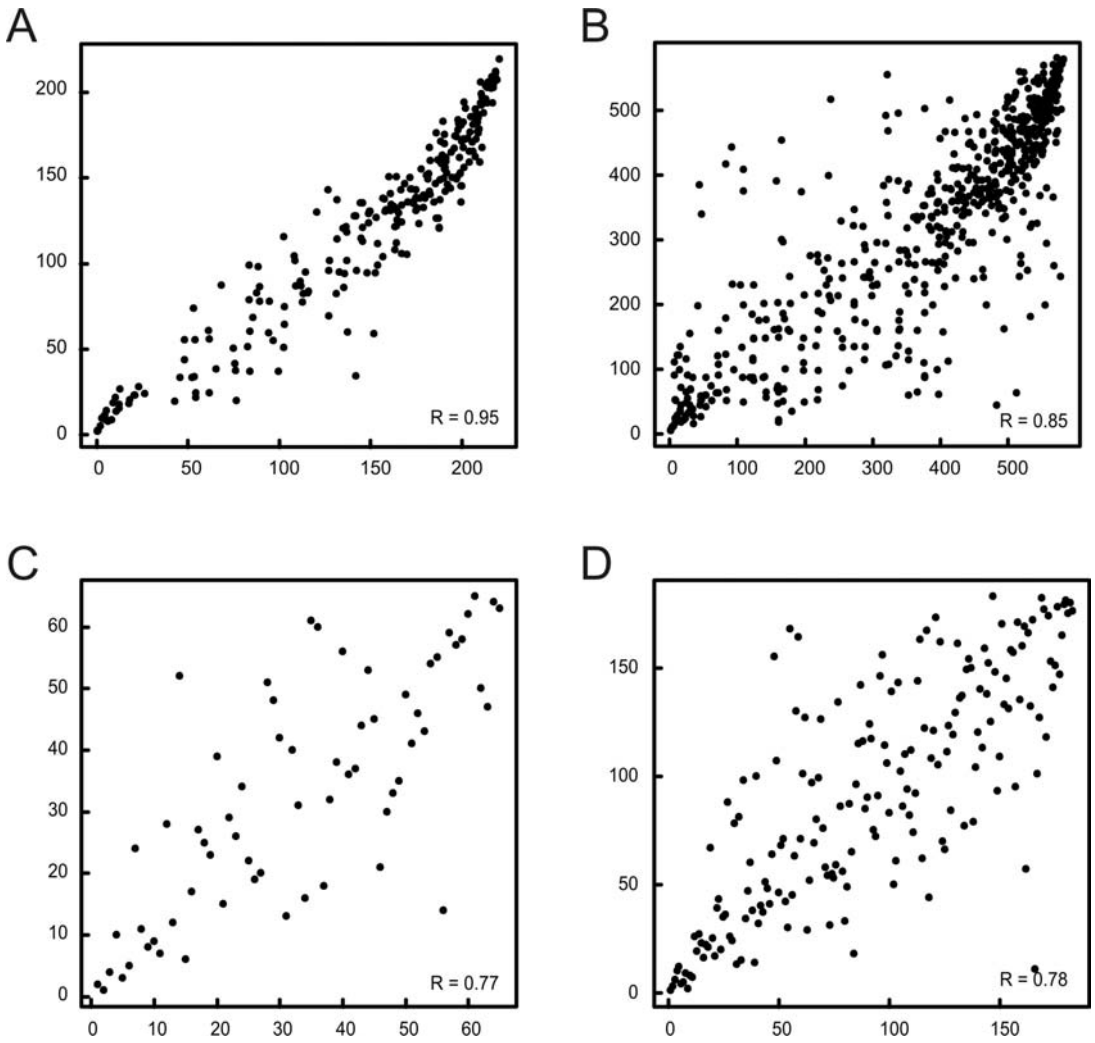


Figure 9.9 – Rank correlation comparison for moderately to highly expressed probes
Rank correlation was used as a metric to investigate comparability of hybridization results between the two array versions. In a first step only moderately to highly expressed probes (signal intensity > 500) were used for comparison. This analysis was performed for (**A**) identical oligonucleotides in the Treg data set, (**B**) cross-annotated probes in the Treg data set, (**C**) identical oligonucleotides in the whole blood data set, and (**D**) cross-annotated probes in the whole blood data set.

In a second step we used probes called present in either one of the sub-groups. Performing rank correlation calculations within the T_{reg} data set, we observed a rank correlation of 0.84 for the identical oligonucleotides and a rank correlation of only 0.69 for the cross-annotated probes (**Figure 9.10A**, **B**). When performing the comparison within the peripheral blood data set, the rank correlations dropped to 0.66 for the identical oligonucleotides and to only 0.55 for the cross-annotated probes (**Figure 9.10C**, **D**).

The strong decrease in rank correlation within the whole blood data set is most likely due to a significant decrease in signal intensities of single probes on the I-huBC-V2 array resulting in large rank differences. To prove this postulate we calculated differentially expressed genes (FC > 1.75, p-value

<0.05, difference of means > 100) between scleroderma and bacteremia samples for the I-huBC-V1 array and determined the corresponding signal values for these genes on the I-huBC-V2 array (**Supplementary Table 3**).

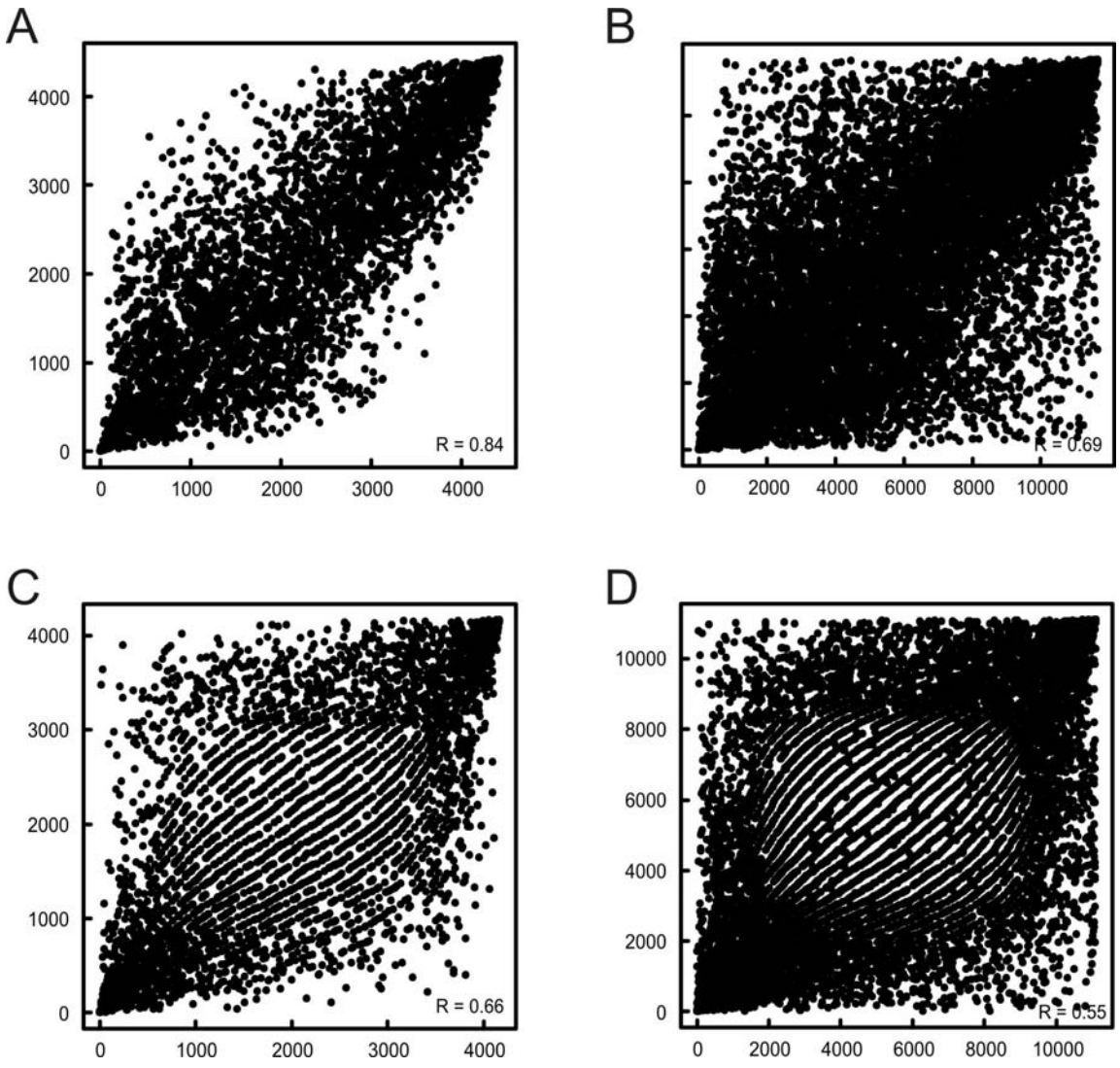


Figure 9.10 – Rank correlation comparison for non-absent probes
Rank correlation was used as a metric to investigate comparability of hybridization results between the two array versions. In the second step all probes which were present in either one of the sub-groups were used. Again, this analysis was performed for (**A**) identical oligonucleotides in the Treg data set, (**B**) cross-annotated probes in the Treg data set, (**C**) identical oligonucleotides in the whole blood data set, and (**D**) cross-annotated probes in the whole blood data set.

Indeed, we detected several probes, which were called differentially expressed on the I-huBC-V1 array, but had very low signal values for both sub-groups on the I-huBC-V2 array. Due to the lower detection limit of the I-huBC-V2 array, these probes were not called absent. To rule out that this difference was intrinsic to the peripheral blood samples we performed the same analysis for the T_{reg}

dataset. Again, several probes were detected that showed significant differences between T_{reg} and T_{conv} cells on the I-huBC-V1 but not on the huBC-V2 array. Similar to the peripheral blood dataset, these probes showed low signal values for both T cell sub-groups on the I-huBC-V2 array (**Supplementary Table 4**). Among these probes was also FOXP3, which is the most important marker of T_{reg} cells. As shown in **Figure 9.6C**, differential expression of FOXP3 between T_{reg} and T_{conv} cells was already confirmed by quantitative RT-PCR as well as intracellular FACS analysis to assess protein expression. Therefore, the data generated with I-huBC- V1 reflected real differences between the tested sub-groups while the I-huBC-V2 did not. BLAST Analysis of the FOXP3 probes revealed distinct yet perfect hits (100 % identity) for both I-huBC-V1 and I-huBC-V2 (data not shown) suggesting that a functional probe was exchanged by a non- functional.

Generalized impact analysis on array performance when upgrading array technologies

To balance the constant need for updates of microarray technologies with the continuation of long-term projects dependent on transcriptional profiling we propose a generalized impact analysis consisting of the *in silico* analysis introduced here combined with an experimental performance analysis as described above (**Figure 9.11**).

The in-silico analysis includes the following steps:

1. Re-blasting of probe sequences from both array types (A and B) using the most up-to-date database annotation.
2. Collecting perfect hits (100% identity) for each probe.
3. Determining the number of hits which are achieved by both array types (“c”) or only by array type A (“a”) or B (“b”), respectively and categorization of hits according to Table 1.

For the subsequent performance analysis individual samples should be hybridized to both array types A and B. The biological samples used for this performance analysis should fulfill the criteria mentioned above.

The experimental analysis includes the following steps:

1. Cross-annotation of data sets generated on arrays to be compared.
2. Sensitivity analysis using determination of absent or present status of probes.
3. Analysis of dynamic range and background values visualized by boxplots.
4. Comparability of signal values using quidproquo technical replication.

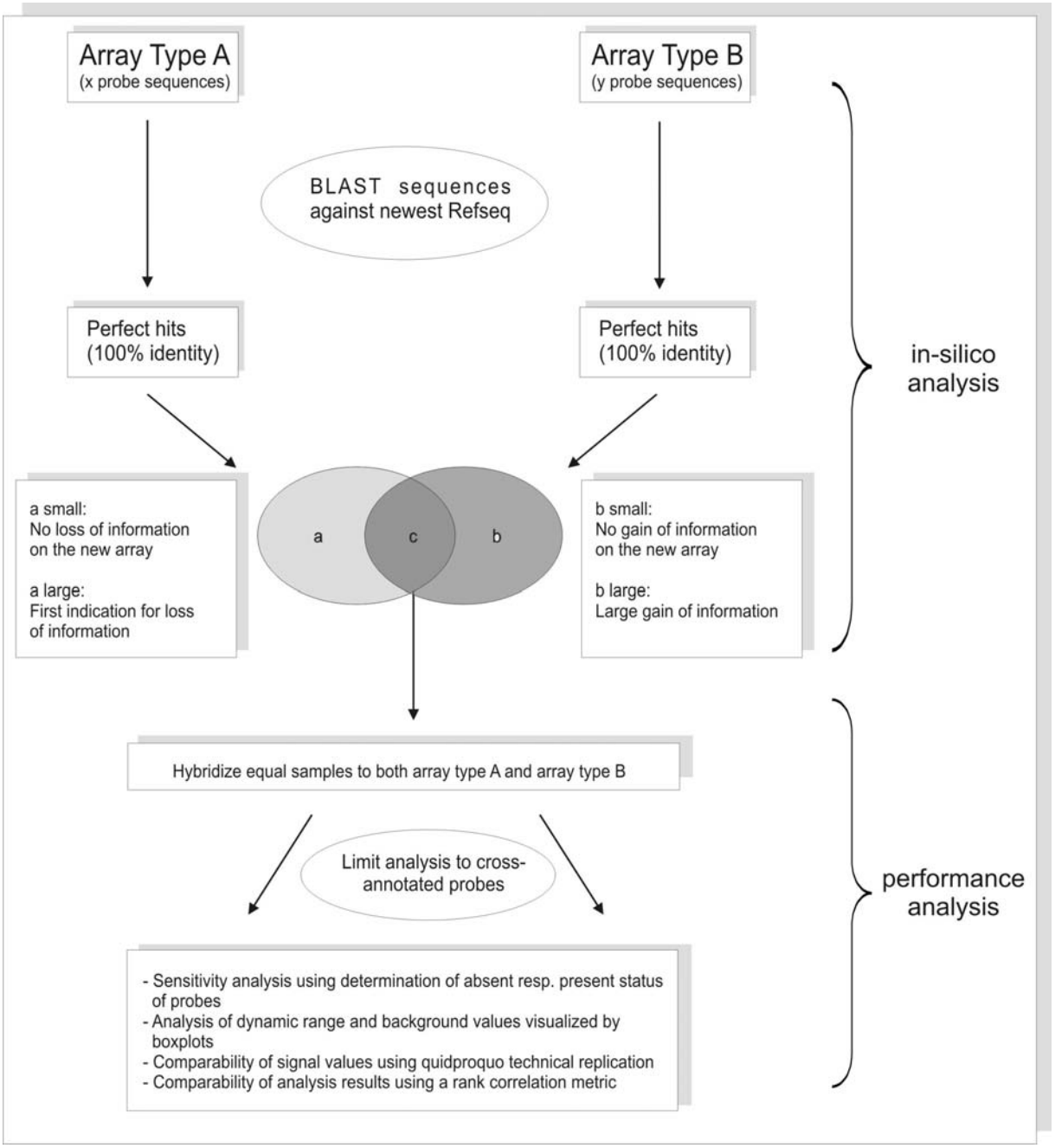


Figure 9.11 – Workflow diagram
Proposed method to quickly determine the impact of changes between subsequent microarray versions. This generalized impact analysis consists of an *in silico* analysis combined with an experimental performance analysis.

9.3 Discussion

Here we have addressed the overall impact of improvements of genomic database content and annotation over time and the impact of technology optimization on major performance issues of a

typical microarray analysis. Unexpectedly, database content and annotation as exemplified for the Refseq database still remains highly dynamic, which by itself has a significant impact on microarray probe annotation. Using an *in silico* approach based on BLAST analysis combined with categorization of probes and respective cross-annotation approaches, we demonstrate that content changes on a given microarray platform are also influenced by database dynamics. Moreover, we conducted a performance analysis combining common quality control measures with a rank correlation metric and show that the inclusion of real biological experiments is mandatory to estimate the overall impact of technology improvements on data consistency. Using the Illumina BeadChip platform as an example, we demonstrate that a large change of probe content between subsequent array versions results in incompatible data in addition to unexpected challenges, such as significant introduction of non-functional probes. This has high impact on biological screening experiments, when signals for known marker genes are lost (as exemplified for FOXP3). Even higher impact can be expected for experiments within a diagnostic setting, where content and technology changes will lead to incompatible diagnostic signatures.

Part IV: Summary and future directions

In 2003, Joseph Nevins' lab introduced an elegant approach for prediction of oncogenic pathway activity in mouse (Huang, Ishida et al. 2003) which was further developed to derive human oncogene-specific gene signatures *in vitro.* Bild and colleagues used this approach, created specific gene signatures for different oncogenes and applied these gene signatures to tumor samples to demonstrate the existence of these molecules *in vivo* (Bild, Yao et al. 2006). Up to then, different types of signatures had been established for diseases and biological processes alike (van de Vijver, He et al. 2002; Baechler, Batliwalla et al. 2003; Yagi, Morimoto et al. 2003; Bertucci, Borie et al. 2004), but this was the first time to define signatures from cell lines *in vitro* and successfully applying these signatures to obtain *in vivo* results. Taking up this approach, we hypothesized that - in principle - this concept should be applicable to any other cell and factor that leads to transcriptional changes upon stimulation and signaling. That means, observed transcriptional changes are biological responses of any given cell in reply to a molecular signal and can therefore be termed a RNA fingerprint of the respective signal. Additionally to molecule specific RNA fingerprints we further hypothesized that generally all predictive gene signatures generated from transcriptional profiles could be considered as RNA fingerprints, including disease specific signatures. Following these hypotheses several different concepts of RNA fingerprints were developed and were introduced in the course of this thesis.

To state the original concept of determining gene signatures *in vitro* and applying them in *vivo* as a concept of RNA fingerprints , we generated gene signatures for different T cell inhibitory molecules, including TGFβ and PD-1, and termed the generated signatures RNA fingerprints of these molecules. These RNA fingerprints should then provide direct evidence whether the cells within a tumor environment are under the control of the interrogated molecules. Using supervised and unsupervised classification methods based on the RNA fingerprints of both, TGFβ and PD-1 we were able to show that T cells derived from patients with Hodgkin's lymphoma are indeed under the influence of both, TGFβ and PD-1. When interrogating T cells from patients with follicular lymphoma, no influence of either TGFβ or PD-1 could be determined. This study was a starting point for RNA fingerprints which clearly demonstrated that this concept can be used to establish RNA fingerprints of diverse signaling molecules *in vitro* and testing them *in vivo*. Next steps include the interrogation of other molecules for which the direct evidence of contribution *in vivo* is still unclear. In terms of the already mentioned inhibitory molecules within a tumor environment, the next step has already been initiated. In addition to the already established fingerprints of PD-1 and TGFβ, fingerprints for CTLA4,

PGE_2 (Chemnitz, Driesen et al. 2006), VEGF and IL10 were generated. The fingerprints of these different inhibitory molecules on the one hand share common features, but are on the other hand also quite distinct from each other. These common and distinct features can now be used to numerically quantify the inhibitory effect of the different molecules, but also to distinguish each RNA fingerprint from each other by identifying specific inhibitory features for each of these molecules. This will hopefully lead to further functional characterization of the signaling pathways these molecules are involved in.

The second concept introduced the use of a disease specific RNA fingerprint in a diagnostic setting. Here we biologically defined an RNA fingerprint and further used it to predict the occurrence of lung cancer prior to clinical manifestation. For the determination of a lung-cancer specific RNA fingerprint we postulated that there are specific genome-wide transcriptional changes in peripheral blood from patients with clinical manifest lung cancer compared to control patients which can be used as an RNA fingerprint for this disease. We furthermore hypothesized that this fingerprint is an early event in lung cancer development and might therefore be suitable for early detection of lung cancer. Indeed we demonstrated differential expression of several genes between patients with lung cancer and controls and showed that the generated RNA fingerprint can be used to detect individuals developing lung cancer prior to clinical manifestation. As already stated in Chapter 5.4, this study was a prospective study using very few samples which provided a first hint on whether lung cancer can be predicted prior to clinical manifestation. To substantiate these findings, the study has to be repeated within a larger setting, composing at least 200 patients with lung cancer and 200 controls for the creation of the RNA fingerprint. Additionally, the prospectively observed cohort has to be enlarged to analyze another 200-400 samples. We have already started to substantiate the predictive ability of the RNA fingerprint on a validation cohort composed of another 37 samples including 22 patients with manifest lung cancer.

The third concept dealt with the use of pre-defined RNA fingerprints. In contrast to the preceding concepts where the RNA fingerprint was determined by analyzing biological experiments, another aspect would be to use pre-defined RNA fingerprints and test these fingerprints for their contribution in separating interrogated subgroups in a microarray experiment. This concept then resembles a gene-class testing approach and the pre-defined fingerprints used for this purpose can be extracted from biological databases which include information about genes belonging to special pathways or groups of genes with similar functions. Different algorithms have been introduced for gene-class testing (Khatri, Bhavsar et al. 2004; Khatri, Desai et al. 2006) with Gene set enrichment analysis (GSEA) as the “gold-standard” (Mootha, Lindgren et al. 2003; Patti, Butte et al. 2003; Petersen,

Dufour et al. 2004; Subramanian, Tamayo et al. 2005). I have developed a new and very simple method, GOAna, which is based on RNA fingerprints provided by the Gene Ontology (GO) Consortium (Ashburner, Ball et al. 2000) and implements such a gene-class testing approach. Using GOAna, it is possible to perform an unbiased analysis based on all branches of GO. Testing a new algorithm always includes the comparison to the mostly applied method, in this case GSEA. Using the original data set Mootha and colleagues initially performed GSEA on (Subramanian, Tamayo et al. 2005), GOAna was compared to GSEA. Both methods obtained differing results which was mainly explained by the kind of approach taken. While GOAna was carried out using all processes included in GO, Mootha and colleagues used 149 hand-curated gene sets for their analysis. The most significant gene spaces obtained by GOAna were not included in these gene sets. Despite these differences, when restricting the analysis to the 149 gene sets used by Mootha and colleagues, GOAna obtained the same results as GSEA which qualifies GOAna as an easy-to-use gene-class testing approach for unbiased analysis of microarray experiments. Further developments of the algorithm are already planned. To date, GOAna is implemented as an R-package which, for computationally extensive algorithms, is not the most well suited programming language. The effort goes towards implementing the algorithm within a JAVA or C++ environment and additionally adding a graphical user interface. These further developments could be part of a Diploma or Master student's thesis. On the other hand, GOAna could be included in the software project "IlluminaGUI" which to date implements a graphical user interface for diverse data analysis methods for gene expression data from the Illumina platform (see Chapter 8). IlluminaGUI was intended to enable the interested life scientist who is not familiar with a command line based environment like the R language to analyze microarray experiments. Besides the already mentioned extension of IlluminaGUI towards GOAna it is planned to add diverse features for analysis of high throughput miRNA data.

The fourth concept introduced the idea of using the microarray experiment itself as a RNA fingerprint. We hypothesized that all transcriptional changes which are revealed by a microarray experiment can serve as a RNA fingerprint and can decipher underlying signaling mechanisms. A large amount of regulatory and signaling mechanisms are not happening on the transcriptional, but on post-transcriptional/protein level. Reversible phosphorylation of proteins, for example, is an important regulatory mechanism. Enzymes called kinases (phosphorylation) and phosphatases (dephosphorylation) are involved in this process. Many enzymes and receptors are switched "on" or "off" by phosphorylation and dephosphorylation. These mechanisms cannot be directly interrogated by transcriptional profiling methods, but they introduce transcriptional changes which can then indirectly be analyzed using a microarray experiment. In the original GOAna algorithm, we analyzed transcriptional profiles in an unbiased way to determine processes involved in the separation of the

examined subgroups. To further analyze the identified processes, the original GOAna algorithm was extended by a network construction step to derive a network of contributing genes, i.e. genes which appear several times in the gene sets called significant by the gene class testing step. We hypothesized that these genes were likely to act as key players in underlying signaling events by linking the different processes which were identified as separating the examined subgroups. Using the extended algorithm we compared activated $CD4^+$ T cells to activated $CD4^+$ T cells in the presence of PGE_2 to get a hint for the signaling events occurring in an inhibitory environment introduced by PGE_2. We identified PP2A as the most prominent gene included in the most significant gene spaces and therefore linking these gene spaces. PP2A is a known phosphatase which has been described as a central regulator in diverse signaling pathways. To interrogate whether PP2A is indeed involved in PGE_2 signaling we subsequently analyzed the regulatory ability of PP2A in the presence or absence of PGE_2 using a Jurkat cell line. In a Western blot analysis we found that PGE_2 acts as a repressor on PP2A, at least in the Jurkat cell line. Many of the fundamental insights into T cell receptor signaling came from studies carried out with transformed T cell lines, especially the Jurkat cell line. For example, using Jurkat cell lines it was elucidated that TCR signaling works through protein tyrosine kinase signaling. Other findings included insights into calcium signaling (Abraham and Weiss 2004). However, there are several problems associated with the use of Jurkat cell lines. Compared to primary T cells, Jurkat cells were shown to be defective in the expression of the lipid phosphatase PTEN (phosphatase and tensin homologue). When PTEN is absent, an important signaling pathway, the PI3K-signalling pathway, is constitutively activated. It is still unclear to what extent the abnormal PTEN status of Jurkat cells alters their response to TCR stimulation and also the status of comparability of primary T cells and the Jurkat cell line. The next step in this project is therefore to perform the same experiment in primary $CD4^+$ T cells. This will either confirm the findings obtained in the Jurkat cell line or conquer these findings. In the case of differing findings the difference of Jurkat cell lines and primary T cells has to be taken into account.

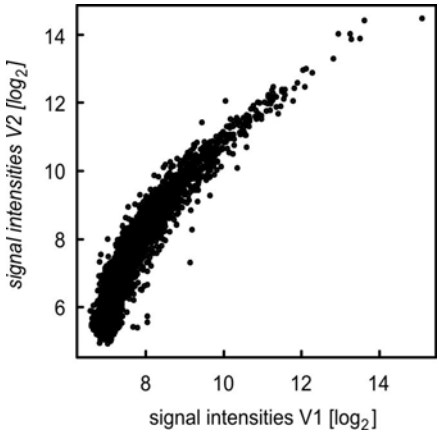
All described approaches for the creation of RNA fingerprints are heavily dependent on the reliability of the underlying technology, in this case the microarray format used for the study. Reproducibility of RNA fingerprints on different platforms or the continuity of RNA fingerprints when new generations of microarrays with updated probe content become available are only two considerations. Concerning the reproducibility of RNA fingerprints, most recently, validity and comparability of transcriptional profiling using different microarray platforms has been very elegantly demonstrated by the MAQC consortium (Canales, Luo et al. 2006; Guo, Lobenhofer et al. 2006; Patterson, Lobenhofer et al. 2006; Shi, Reid et al. 2006; Shippy, Fulmer-Smentek et al. 2006; Tong, Lucas et al. 2006). The MAQC project clearly demonstrated that comparability of microarray technology is

already high 1) when restricting the analysis to a comparable set of data points (genes) and 2) when comparing high throughput technologies developed approximately at the same time. Proving consistency of these technologies when introducing technological advances, i.e. the analysis of new microarrays with updated probe content was suggested by MAQC as a major issue for future development. Indeed, when interrogating improvements of genomic database content and annotation over time we saw that, unexpectedly, database content and annotation still remain highly dynamic. This by itself has a significant impact on microarray probe annotation. Therefore, we have embarked on the major task of comparing subsequent generations of microarrays and have developed a methodology to rapidly determine the impact of probe changes on reproducibility and comparability of microarray results. Using the Illumina BeadChip platform as an example, we demonstrated that a large change of probe content between subsequent array versions results in incompatible data in addition to unexpected challenges, such as significant introduction of non-functional probes. This has high impact on biological screening experiments, when signals for known marker genes are lost (as exemplified for FOXP3). Even higher impact can be expected for experiments within a diagnostic setting, where content and technology changes will lead to incompatible diagnostic signatures. A next important step in genomic sciences would therefore be to quickly introduce standardized general impact analyses to assess newer generation technologies. It would be desirable to introduce the presented approach as a starting point for further projects within the MAQC consortium. Next steps could be to test the overall impact of the presented approach in the larger consortium and perform such impact analyses on a grand scale respectively when new technologies become available again. There is still the question of what influence the presented difficulties in comparability and reproducibility of results imply on the concept of RNA fingerprints. The concept or RNA fingerprints which is called like this because the fingerprints are derived from transcriptional profiling studies is left with an aftertaste. If there are fingerprints developed for diverse signaling molecules or even disease specific fingerprints, we would like to be able to recall and reuse these fingerprints even if a new generation of microarrays is distributed. One example of a collapsing RNA fingerprint is the lung cancer specific fingerprint presented in Chapter 5. During this study Illumina distributed the I-huBC-V2 array presented in Chapter 9 and we repeated the study on this new array platform. Strikingly, since only 8299 probes were identical on both platforms (see Chapter 9) only 74 of our 154 genes included in the RNA fingerprint could be recalled. The RNA fingerprint could no longer be used as it was. Therefore a new fingerprint had to be calculated based on the new array format. We were still able to achieve a new RNA fingerprint with predictive ability, but when going into a clinical setting these difficulties have to be eliminated to derive true disease specific RNA fingerprints which are predictive for the disease. On the other hand

the different concept of RNA fingerprints presented in this thesis are still usable, since these are concepts which can be used with any underlying technology which measures the abundance of transcripts. With emerging technologies on the microarray side (Hardiman 2006; Hoheisel 2006; Shi, Reid et al. 2006) and the use of methodologies like the one introduced in Chapter 9, these concepts can be used without concern. The presented methodology will help researchers to evaluate whether an established microarray format should be continuously used (see in the case of lung cancer specific fingerprint) or whether a change to a newer generation can be carried out without damage of the RNA fingerprint. A further new emerging technology is high throughput sequencing (Bentley 2006; Kim, Porreca et al. 2007; Velculescu and Kinzler 2007). With this technology the concepts of RNA fingerprints can be used without any concerns, since RNA abundance is digitally quantified which avoids any probe sequence or annotation difficulties.

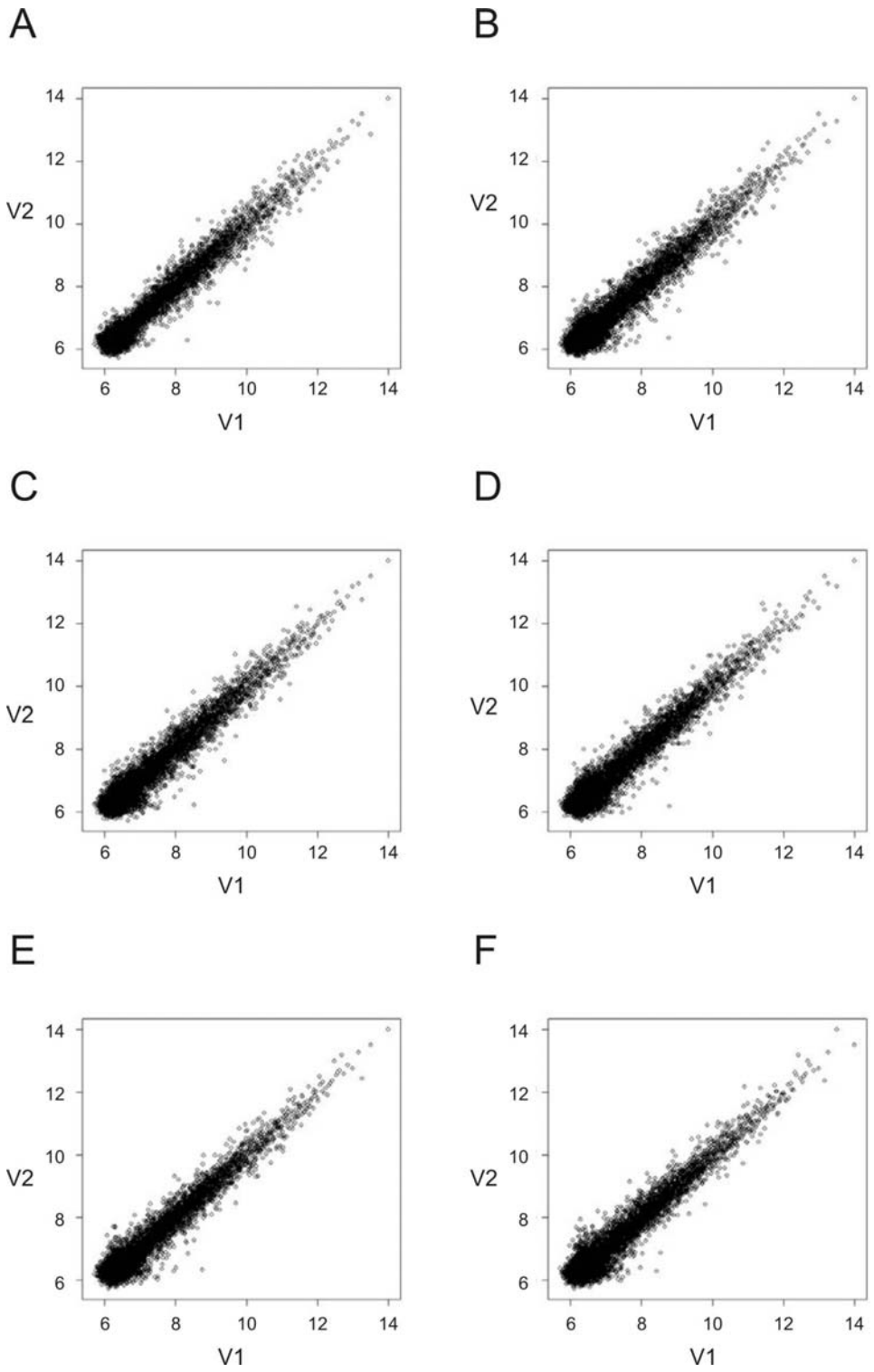
Appendix A – Supplementary Figures

Supplementary Figure 1



Supplementary Figure 1 – Comparison of signal intensities between subsequent array versions
Raw signal intensities for quidproquo technical replicates were compared using pairwise scatterplots. Depicted is one example. log_2 signal values of the 8299 identical probes from I-huBC-V1 and I-huBC-V2are plotted on the x-axis and y-axis, respectively.

Supplementary Figure 2



Supplementary Figure 2 – Correlation of technical replicates in the T_{reg} data set
To investigate the outcome of technical replication we used pairwise scatterplots. For perfect technical replicates one would expect a straight diagonal line in a pairwise scatterplot. Data for both array versions was limited to 8,299 identical oligonucleotides. Pairwise scatterplots of signal intensities were performed on the normalized T_{reg} set. Shown are scatterplots for samples 1-6 (**A-F**).

Appendix B – Supplementary Tables

Supplementary Table 1

Comparison	Correlation identical oligos	Correlation cross-annotated probes
1_V1 vs 1_V2	0.97	0.69
2_V1 vs2_V2	0.97	0.66
3_V1 vs 3_V2	0.97	0.65
4_V1 vs 4_V2	0.97	0.67
5_V1 vs 5_V2	0.96	0.65
6_V1 vs 6_V2	0.96	0.7

Supplementary Table 1 – Pairwise correlations in the Treg data set
Depicted are pairwise correlations of technical replicates in the T_{reg} data set. Correlations were calculated using Pearson's correlation coefficient implemented in R (R Development Core Team 2007).

Supplementary Table 2

Comparison	Correlation identical oligos	Correlation cross-annotated probes
1_V1 vs 1_V2	0.98	0.54
2_V1 vs2_V2	0.96	0.62
3_V1 vs 3_V2	0.99	0.55
4_V1 vs 4_V2	0.96	0.58
5_V1 vs 5_V2	0.98	0.64
6_V1 vs 6_V2	0.94	0.62
7_V1 vs 7_V2	0.98	0.68
8_V1 vs 8_V2	0.96	0.64
9_V1 vs 9_V2	0.96	0.71
10_V1 vs 10_V2	0.96	0.58
11_V1 vs 11_V2	0.97	0.56
12_V1 vs 12_V2	0.97	0.72
13_V1 vs 13_V2	0.94	0.69
14_V1 vs 14_V2	0.98	0.63
15_V1 vs 15_V2	0.98	0.61
16_V1 vs 16_V2	0.96	0.71

Supplementary Table 2 – Pairwise correlations in the peripheral blood data set
Depicted are pairwise correlations of technical replicates in the T_{reg} data set. Correlations were calculated using Pearson's correlation coefficient implemented in R (R Development Core Team 2007).

Supplementary Table 3

Probeset	Symbol	mean Sclero V1	mean Bact V1	FC	pval	diff	mean Sclero V2	mean Bact V2	FC	pval	diff
5220477	IFI27	634.5	2296	-3.57	0.40	1661	301.1	1416	-4.76	0.37	1115
4120707	RPL23	424.5	1251	-2.94	0.00	827	588.2	1501	-2.56	0.01	912
7000608	RPS7	553.8	1646	-2.94	0.00	1092	60.09	70.7	-1.18	0.24	11
2490056	LOC644972	1988	5276	-2.63	0.00	3288	1176	3199	-2.70	0.00	2023
3310091	DEFA3	1196	3089	-2.56	0.09	1892	2870	6073	-2.13	0.07	3203
2000025	RPL26	276.7	693.6	-2.50	0.05	417	428.1	1353	-3.13	0.02	925
6960440	DEFA4	306.4	766.8	-2.50	0.08	460	276.3	726.7	-2.63	0.16	450
3290605	TOMM7	350.8	853.3	-2.44	0.00	502	1573	3219	-2.04	0.00	1645
7610544	LOC644790	1727	4000	-2.33	0.00	2273	73.39	103.1	-1.41	0.03	30
130215	RPL39	2112	4949	-2.33	0.00	2837	2420	5211	-2.17	0.00	2791
1820056	COX6C	241.8	558	-2.33	0.01	316	294.1	681	-2.33	0.02	387
70722	COX7B	221.1	500.4	-2.27	0.01	279	97.64	247	-2.50	0.00	149
1340192	C15orf15	229.7	510	-2.22	0.00	280	174.5	424	-2.44	0.00	250
6370367	GZMA	457.4	1026	-2.22	0.01	569	977	2066	-2.13	0.01	1089
1450390	KLRB1	243.5	545.4	-2.22	0.00	302	165.5	355	-2.13	0.00	189
1260278	RPL41	6009	13460	-2.22	0.00	7451	56.34	64.29	-1.14	0.02	8
7330129	HINT1	682	1474	-2.17	0.00	792	632.7	1381	-2.17	0.00	748
4250035	RAP1GAP	227.5	488.7	-2.13	0.43	261	324.8	781.4	-2.38	0.44	457
430328	ERAF	2433	5212	-2.13	0.14	2779	2909	5279	-1.82	0.21	2370
6960554	LCN2	392.3	811.9	-2.08	0.01	420	495.9	1026	-2.08	0.01	531
4730612	RPS17	496.4	1031	-2.08	0.03	535	861.6	2015	-2.33	0.01	1154
2940639	RPL9	328.8	661.5	-2.00	0.02	333	1767	4511	-2.56	0.00	2744
520161	TPT1	1241	2497	-2.00	0.00	1256	4752	8064	-1.69	0.00	3312
5310369	GZMK	264	525.5	-2.00	0.00	261	553.5	1255	-2.27	0.00	702
6980474	LY96	314.1	614.5	-1.96	0.00	300	369.9	788.9	-2.13	0.00	419
6280576	S100A8	4923	9574	-1.96	0.00	4651	9568	15306	-1.59	0.00	5738
6660220	RPS10	2734	5408	-1.96	0.00	2674	7494	11017	-1.47	0.00	3523
4070164	COX7C	1136	2226	-1.96	0.01	1090	774.4	1815	-2.33	0.00	1040
7320709	NA	5495	10820	-1.96	0.00	5325	2959	5382	-1.82	0.01	2423
5290523	NA	5495	10820	-1.96	0.00	5325	8808	16142	-1.82	0.00	7334
3190193	RPL30	1190	2310	-1.92	0.00	1120	5137	7657	-1.49	0.00	2520
1990634	TIMM8B	189	366.6	-1.92	0.00	178	71.97	91.75	-1.28	0.04	20
3400551	MS4A3	127.4	246.5	-1.92	0.08	119	94.32	257.6	-2.70	0.09	163
1690605	RPS27L	185.9	353.7	-1.89	0.02	168	191.1	479.9	-2.50	0.03	289
160348	RNASE3	168.8	316.6	-1.89	0.06	148	173.4	415.8	-2.38	0.07	242
5670601	RPL35A	2074	3926	-1.89	0.00	1851	2809	4564	-1.61	0.00	1755
2900593	PFDN5	1025	1950	-1.89	0.02	925	1785	3401	-1.92	0.01	1616
270451	NDUFA4	769.9	1435	-1.85	0.02	665	923.7	1804	-1.96	0.00	880
2320403	RPL27	1857	3461	-1.85	0.02	1604	3767	6282	-1.67	0.00	2515
6370181	RPL11	1605	2941	-1.82	0.01	1337	3745	6320	-1.69	0.00	2574

6400736	CAMP	570.6	1039	-1.82	0.02	469	923.3	1578	-1.69	0.03	655
6770246	EEF1A1	5356	9594	-1.79	0.00	4237	5878	7438	-1.27	0.10	1560
2190669	NA	127.2	225.8	-1.79	0.03	99	42.4	43.71	-1.03	0.66	1
1170400	C12orf57	315.8	567	-1.79	0.01	251	786.4	1309	-1.67	0.04	523
3190053	SNRPD2	717.4	1280	-1.79	0.01	563	538.2	1004	-1.85	0.01	466
2900356	SCIN	6780	3827	1.77	0.00	2952	45.31	46.6	-1.03	0.25	1
4390692	HLA-DRB5	1500	849.7	1.77	0.13	651	1126	303.7	3.71	0.04	823
7330093	HLA-DRB1	371.1	208	1.78	0.45	163	1139	366.7	3.11	0.12	772
4490017	NFATC3	1708	958.9	1.78	0.00	749	48.38	51.86	-1.08	0.50	3
3060487	LOC255374	1658	928.1	1.79	0.00	730	45.7	47.11	-1.03	0.51	1
6370035	OASL	436.5	242.6	1.8	0.12	194	832.3	372.9	2.23	0.08	459
3400142	TADA3L	905.3	500.3	1.81	0.00	405	46.44	45.41	1.02	0.38	1
5890196	IL18	2740	1502	1.82	0.00	1238	11716	8570	1.37	0.00	3146
1400722	TPPP3	264.1	143.3	1.84	0.00	121	82.03	53.68	1.53	0.00	28
2690452	IFIT3	765.4	406.2	1.88	0.16	359	2377	931.8	2.55	0.09	1445
6200376	IFITM2	2629	1392	1.89	0.00	1238	17389	14419	1.21	0.04	2970
3310725	PARP10	981.9	517.2	1.9	0.01	465	899.6	557.6	1.61	0.02	342
4220435	OAS3	433.9	220.7	1.97	0.10	213	474.3	190.6	2.49	0.06	284
3850524	CEP27	920.4	465.4	1.98	0.00	455	1262	733	1.72	0.00	529
5870047	NA	1630	815.5	2	0.04	814	2458	1353	1.82	0.06	1104
6620711	RSAD2	714.6	338.9	2.11	0.04	376	610.3	181.5	3.36	0.07	429
4060674	IL1RN	698.5	325.1	2.15	0.00	373	905.5	525.1	1.72	0.02	380
630450	IFIT2	1373	631.6	2.17	0.04	742	2990	1403	2.13	0.05	1587
1780632	IFIT1	553.3	250.2	2.21	0.07	303	1608	555.7	2.89	0.06	1053
2630110	MX1	1888	847.8	2.23	0.06	1040	6394	2751	2.32	0.04	3643
7200255	IFI44L	923.7	409.5	2.26	0.15	514	1502	596.3	2.52	0.12	906
2230601	FOLR3	1005	430.3	2.34	0.13	575	2486	1044	2.38	0.09	1443
1710259	HERC5	1193	503.6	2.37	0.08	690	2445	987.2	2.48	0.08	1458
1070528	ISG15	2733	1076	2.54	0.11	1658	2783	1037	2.68	0.10	1746
4280725	HES4	485.6	182.4	2.66	0.07	303	819.3	266.3	3.08	0.05	553

Supplementary Table 3 – Differentially expressed genes in the peripheral blood data set
Differentially expressed genes (FC > 1.75, p-value < 0.05, diff > 100) between Scleroderma and Bacteremia samples on I-huBC-V1 were calculated and corresponding values for these genes were checked on I-huBC-V2. Marked in grey are genes which show very low signal intensities on I-huBC-V2.

Supplementary Table 4

Probeset	Symbol	mean Tregs V1	mean non-Tregs V1	FC	pval	diff	mean Tregs V2	mean non-Tregs V2	FC	pval	diff
2260148	NELL2	117	617	-5.27	0.00	500	87	1118	-12.88	0.00	1032
7380181	LRRN3	133	670	-5.04	0.00	537	84	1048	-12.41	0.00	963
1990446	CCL5	298	1257	-4.22	0.00	959	134	492	-3.67	0.00	358
4670092	ID2	170	594	-3.5	0.00	424	236	1491	-6.32	0.00	1255
5310369	GZMK	294	950	-3.23	0.01	656	700	2670	-3.82	0.00	1970
6760075	EOMES	93	285	-3.05	0.00	192	72	520	-7.18	0.00	448
6860129	PCSK5	141	401	-2.84	0.00	260	97	323	-3.32	0.00	226
4670056	ANKRD55	112	282	-2.52	0.00	170	113	465	-4.11	0.00	352
7570079	IL7R	673	1668	-2.48	0.00	995	735	1812	-2.46	0.00	1076
2230152	SATB1	499	1240	-2.48	0.00	741	141	490	-3.47	0.00	349
4010301	C1orf162	481	1133	-2.35	0.00	651	1355	3257	-2.4	0.03	1902
2120500	ACTN1	386	901	-2.33	0.00	515	472	1036	-2.19	0.00	564
4850128	GZMH	97	225	-2.31	0.01	127	56	385	-6.87	0.01	329
6620279	C6orf190	140	319	-2.27	0.00	179	128	338	-2.65	0.00	210
5810392	BHLHB2	142	317	-2.23	0.00	175	139	541	-3.9	0.01	403
6400242	TMEM71	378	822	-2.17	0.00	443	633	1605	-2.53	0.00	972
2710239	LASS6	136	292	-2.15	0.00	156	188	651	-3.46	0.00	463
1770332	NA	157	334	-2.13	0.00	177	73	105	-1.44	0.00	32
5220259	TARP	153	323	-2.12	0.00	171	87	183	-2.09	0.00	96
830484	AIF1	150	315	-2.1	0.00	165	207	677	-3.27	0.00	470
4480367	KRT72	119	246	-2.07	0.07	127	118	1176	-9.97	0.08	1058
6270128	CD40LG	147	297	-2.03	0.00	151	134	440	-3.29	0.00	306
2470292	MAN1C1	165	334	-2.02	0.00	169	227	492	-2.17	0.01	265
4540424	EPHX2	161	319	-1.98	0.00	158	345	678	-1.96	0.00	333
840427	SLC40A1	173	338	-1.95	0.00	164	60	68	-1.13	0.10	8
6200397	OXNAD1	218	421	-1.93	0.00	203	652	1637	-2.51	0.00	986
1660403	ANK3	100	194	-1.93	0.00	94	71	229	-3.24	0.00	158
6590561	CCR7	1261	2419	-1.92	0.01	1158	3513	6266	-1.78	0.07	2753
2900463	GNLY	135	257	-1.9	0.05	122	63	372	-5.87	0.02	309
2600706	RAB6IP1	119	225	-1.88	0.00	106	193	708	-3.66	0.00	515
6370367	GZMA	376	708	-1.88	0.21	332	882	1622	-1.84	0.28	740
6180427	GPR160	145	269	-1.86	0.00	125	64	116	-1.83	0.00	53
3460397	FLOT1	436	808	-1.85	0.00	372	368	459	-1.25	0.61	90
5820397	VIPR1	235	436	-1.85	0.00	200	203	398	-1.96	0.00	195
2190678	IL4R	922	1698	-1.84	0.00	776	1839	3415	-1.86	0.00	1576
840253	PLAC8	442	803	-1.82	0.02	361	802	1483	-1.85	0.09	681
670132	RNF144	196	353	-1.8	0.00	157	272	635	-2.34	0.00	363
5820465	PLXDC1	114	203	-1.78	0.00	89	89	327	-3.66	0.00	238
1450400	CHI3L2	135	238	-1.76	0.00	103	107	327	-3.06	0.00	220
4280440	HERPUD1	236	134	1.76	0.00	102	725	240	3.02	0.01	485
4920315	DUSP10	207	118	1.76	0.00	90	233	69	3.37	0.00	164

1340092	ATP2B1	197	112	1.76	0.00	85	133	73	1.82	0.00	60
1850465	SEC11C	519	295	1.76	0.02	224	1216	680	1.79	0.05	536
7160520	C6orf129	382	217	1.76	0.10	165	774	308	2.51	0.18	466
150296	RBX1	1264	718	1.76	0.12	546	2075	1101	1.88	0.17	973
2570564	KLF13	1276	726	1.76	0.14	549	2818	1612	1.75	0.17	1206
6760048	MT2A	920	524	1.76	0.33	396	1401	715	1.96	0.35	685
1170092	TBK1	346	195	1.77	0.00	151	486	180	2.7	0.05	306
2060196	AQP3	342	194	1.77	0.13	148	467	214	2.19	0.18	254
2140598	TUBB2C	582	326	1.78	0.00	256	787	305	2.58	0.00	482
2750528	NA	194	109	1.78	0.00	85	979	135	7.24	0.00	844
1340170	RFTN1	371	208	1.78	0.00	163	1652	727	2.27	0.00	925
2350647	SYT11	252	142	1.78	0.01	111	349	119	2.94	0.05	231
3840019	ARHGAP25	790	444	1.78	0.09	346	2139	1078	1.98	0.16	1060
2810110	TM9SF2	623	350	1.78	0.09	274	926	479	1.93	0.15	447
1170356	PCQAP	291	163	1.79	0.00	128	326	140	2.33	0.11	186
5690477	EMP3	1947	1087	1.79	0.01	860	5310	2968	1.79	0.05	2342
3170598	TULP4	349	194	1.8	0.00	155	571	213	2.68	0.05	358
5570010	NPC1	301	167	1.8	0.01	134	324	140	2.31	0.08	184
5560577	TMPIT	286	158	1.8	0.05	127	566	206	2.75	0.08	360
830047	TXNL5	643	357	1.8	0.06	286	1432	662	2.16	0.08	770
3520072	P4HB	729	404	1.8	0.15	325	810	506	1.6	0.27	304
6100450	SKAP2	211	116	1.81	0.00	95	284	91	3.13	0.00	193
830070	PAM	350	194	1.81	0.00	156	287	136	2.11	0.02	150
2650678	C2orf24	369	204	1.81	0.09	166	652	274	2.38	0.18	378
5870280	NONO	912	504	1.81	0.22	408	206	135	1.52	0.19	71
5860300	HSPA1A	317	174	1.82	0.14	143	1270	372	3.42	0.16	898
4280253	C9orf19	1153	629	1.83	0.00	524	2792	1440	1.94	0.01	1352
6380088	ADCY3	419	227	1.84	0.00	192	929	394	2.36	0.00	535
5560328	PRDX1	2150	1170	1.84	0.00	980	6232	3393	1.84	0.03	2839
7050075	NT5C	276	150	1.84	0.06	126	1886	1058	1.78	0.05	828
1070435	B4GALT3	421	229	1.84	0.10	192	1164	790	1.47	0.16	374
2070634	CERK	898	488	1.84	0.11	410	2199	1015	2.17	0.18	1183
5890139	PRNP	1004	541	1.85	0.00	463	1443	733	1.97	0.03	710
3420274	TK1	186	101	1.85	0.00	85	132	54	2.47	0.02	79
2750730	TRIB2	1076	583	1.85	0.24	493	1704	762	2.24	0.28	942
4150253	FAM110A	536	289	1.86	0.00	247	926	422	2.19	0.01	503
5870743	HLA-DMB	331	178	1.86	0.00	153	591	216	2.74	0.00	375
3850131	LRIG1	332	178	1.86	0.02	154	722	261	2.76	0.08	461
7380671	TBCB	591	318	1.86	0.09	273	1793	843	2.13	0.15	950
670196	CBX7	508	273	1.86	0.09	235	1646	698	2.36	0.17	948
610273	OPTN	1254	671	1.87	0.00	583	820	450	1.82	0.12	369
5690553	NDRG1	764	409	1.87	0.00	355	1051	476	2.21	0.00	575
1110273	SLFN5	992	530	1.87	0.13	461	547	397	1.38	0.14	150
3460224	SLC1A5	364	194	1.88	0.00	170	128	60	2.14	0.08	68
130609	UBL3	574	306	1.88	0.00	268	1001	443	2.26	0.03	558
1580673	GALM	320	170	1.88	0.02	150	621	201	3.09	0.06	420

1710768	TAP1	1480	789	1.88	0.06	691	4583	2582	1.77	0.16	2001
5090184	HEBP2	394	209	1.88	0.08	185	1315	652	2.02	0.09	663
4540048	MCL1	1238	660	1.88	0.16	578	177	154	1.15	0.26	23
1980678	UTS2	244	129	1.89	0.00	115	359	90	3.97	0.00	269
5050487	CCDC23	517	274	1.89	0.02	243	1531	858	1.78	0.08	673
4070133	RNPEPL1	523	277	1.89	0.05	246	170	97	1.75	0.18	73
3840519	BCAS1	185	98	1.89	0.08	87	81	60	1.36	0.37	21
4180382	ADAM8	308	163	1.89	0.10	145	624	209	2.99	0.17	416
6060242	GSTK1	2254	1190	1.89	0.10	1064	7957	4086	1.95	0.17	3870
1770181	KIAA1949	514	273	1.89	0.14	242	2132	1016	2.1	0.12	1116
2940022	LY6E	590	312	1.89	0.26	278	2250	1063	2.12	0.28	1187
240468	ANTXR2	414	218	1.9	0.00	195	605	237	2.56	0.00	368
1500768	DPYSL2	337	177	1.9	0.00	160	857	310	2.76	0.03	547
6180008	RAB37	891	469	1.9	0.00	422	55	54	1.03	0.76	1
6400309	CYBA	1462	770	1.9	0.00	692	4304	2169	1.98	0.06	2135
3940068	SGSH	626	329	1.9	0.07	297	201	180	1.12	0.23	21
5420717	EIF4EBP2	579	305	1.9	0.09	274	1765	828	2.13	0.17	937
2000136	TAGLN2	892	469	1.9	0.12	424	897	579	1.55	0.05	318
1690605	RPS27L	1824	953	1.91	0.01	871	3358	1633	2.06	0.10	1725
3460382	CAPN2	2284	1192	1.92	0.00	1093	208	119	1.74	0.01	89
1240349	TRIM69	210	110	1.92	0.01	101	304	70	4.33	0.07	234
7160164	C6orf108	536	279	1.92	0.10	257	1191	487	2.45	0.15	704
6520497	ACTG1	3611	1873	1.93	0.03	1738	4175	4208	-1.01	0.99	33
1070528	ISG15	1059	549	1.93	0.05	510	1811	731	2.48	0.16	1081
2060497	SLC4A7	558	287	1.94	0.01	271	266	162	1.65	0.00	104
450754	ALDOA	1464	753	1.94	0.24	711	2542	1203	2.11	0.27	1339
2690598	GLB1	287	147	1.95	0.00	140	215	95	2.27	0.01	120
6040008	IQGAP2	482	247	1.95	0.00	235	711	288	2.47	0.03	423
4280129	C16orf24	311	159	1.95	0.01	152	656	208	3.15	0.03	448
3440056	COX8A	1459	749	1.95	0.03	710	4270	2374	1.8	0.12	1896
4200180	AES	735	377	1.95	0.14	357	4438	2146	2.07	0.21	2292
460324	BATF	320	163	1.96	0.00	156	638	250	2.55	0.01	387
5900286	GPR68	262	133	1.96	0.00	128	397	109	3.64	0.04	288
2650079	ACTB	8087	4131	1.96	0.10	3956	6174	5689	1.09	0.84	485
2060241	NA	928	470	1.97	0.02	458	1623	1054	1.54	0.34	570
5550435	TBC1D4	792	403	1.97	0.06	389	1425	695	2.05	0.14	730
2190360	RHOG	836	424	1.97	0.11	412	2649	1205	2.2	0.11	1443
2710253	FAM38A	1011	510	1.98	0.00	501	2784	1231	2.26	0.01	1553
4070681	CENPM	296	149	1.99	0.00	147	721	234	3.08	0.00	487
3780300	TNIP1	582	292	1.99	0.05	290	922	393	2.34	0.14	529
3450692	TRAF1	584	292	2	0.00	291	318	167	1.91	0.01	152
3140280	RCSD1	1209	603	2	0.00	605	779	364	2.14	0.01	415
3180452	PARP12	493	246	2	0.00	247	1477	601	2.46	0.05	877
6560452	JARID1D	444	223	2	0.55	222	438	199	2.19	0.57	238
5570270	CCR5	360	178	2.03	0.01	182	49	45	1.08	0.39	4
4230021	CCR6	227	112	2.03	0.07	116	857	159	5.4	0.05	698

1090274	FAM53B	478	232	2.06	0.00	246	175	101	1.74	0.13	75
5270487	OGDH	367	178	2.06	0.05	189	446	121	3.69	0.09	325
7160669	APBB1IP	572	276	2.07	0.24	296	1365	485	2.81	0.27	880
5890632	PALM2-AKAP2	267	128	2.08	0.00	139	67	63	1.08	0.13	5
5810709	LOC26010	258	124	2.09	0.00	134	261	80	3.25	0.00	181
7400669	HLA-A	1419	678	2.09	0.01	741	16878	9011	1.87	0.05	7867
6270537	MGST2	236	112	2.11	0.00	124	354	103	3.43	0.00	251
7330093	HLA-DRB1	255	121	2.11	0.42	134	385	96	4.02	0.38	290
2320717	PLP2	1017	479	2.12	0.07	538	369	157	2.35	0.10	212
5810681	ICAM3	1622	764	2.12	0.16	858	6270	4113	1.52	0.25	2157
1710021	CDC20	210	98	2.13	0.00	111	211	67	3.15	0.00	144
520711	GDPD5	391	183	2.13	0.01	208	682	164	4.15	0.01	518
460754	NA	1296	608	2.13	0.07	688	5805	2476	2.34	0.17	3329
3400592	CORO7	412	193	2.13	0.12	219	1334	374	3.56	0.15	960
5720192	LSP1	533	249	2.14	0.08	284	184	127	1.45	0.10	58
840110	NA	533	249	2.14	0.08	284	138	98	1.41	0.01	40
1940082	LMNA	234	109	2.15	0.00	125	435	90	4.85	0.00	345
2370673	CD99	1055	491	2.15	0.11	564	2424	1078	2.25	0.11	1346
5390121	HOXB2	364	170	2.15	0.11	195	527	269	1.96	0.23	258
2900626	HN1	1047	482	2.17	0.00	565	780	399	1.96	0.12	382
5390504	BIRC3	3864	1775	2.18	0.07	2089	192	108	1.78	0.03	84
4150161	CD58	295	135	2.19	0.00	160	445	112	3.98	0.01	333
830762	TXN	1052	476	2.21	0.00	576	2064	753	2.74	0.03	1311
1710259	HERC5	567	255	2.23	0.00	312	1176	385	3.05	0.01	790
6660181	MYO1F	289	130	2.23	0.02	160	556	122	4.57	0.09	434
7650719	FAM26B	394	176	2.24	0.00	218	248	89	2.79	0.02	159
5720647	ADAM19	572	255	2.24	0.01	317	1067	368	2.9	0.04	699
6200273	EPSTI1	675	300	2.25	0.00	376	1287	471	2.73	0.04	816
2970201	C1orf78	342	151	2.27	0.00	191	89	59	1.5	0.02	30
6330672	ANXA5	575	253	2.28	0.00	322	1603	455	3.53	0.00	1149
3310368	ID3	528	230	2.29	0.03	297	843	279	3.02	0.09	564
7160390	GAPDH	1631	711	2.29	0.04	920	2135	937	2.28	0.07	1198
5570035	PPP1CA	1084	472	2.3	0.08	613	1274	404	3.15	0.09	869
3460008	KCNN4	351	152	2.31	0.10	199	919	226	4.06	0.12	693
6560274	VIL2	1690	729	2.32	0.00	961	4812	2126	2.26	0.01	2685
3520475	PTTG2	339	145	2.33	0.00	194	58	55	1.06	0.42	3
3420148	CLDND1	1269	544	2.33	0.01	724	2131	839	2.54	0.03	1293
1400482	CHST7	426	182	2.35	0.00	245	720	187	3.84	0.01	532
6330152	KIAA0101	223	95	2.35	0.00	128	410	63	6.45	0.00	346
6130563	ECGF1	538	229	2.35	0.01	309	67	54	1.24	0.00	13
7160390	GAPDH	2545	1083	2.35	0.03	1462	2135	937	2.28	0.07	1198
3940433	F5	290	123	2.36	0.00	167	547	120	4.54	0.00	426
6650300	PPP1R2P9	305	129	2.37	0.12	176	414	112	3.7	0.11	302
4390113	IL32	3352	1408	2.38	0.00	1944	1664	712	2.34	0.00	952
2680370	HLA-DRA	282	118	2.39	0.00	164	1224	133	9.21	0.00	1091

830762	TXN	1582	661	2.39	0.00	922	2064	753	2.74	0.03	1311
7380056	TP53INP1	756	317	2.39	0.00	439	908	301	3.02	0.01	607
1260181	SELPLG	1433	599	2.39	0.08	833	169	146	1.16	0.20	23
3420356	RGS1	1682	703	2.39	0.27	979	4304	2046	2.1	0.31	2258
50072	SEMA3G	258	107	2.4	0.00	150	141	48	2.93	0.00	93
7050433	CLIC1	1634	674	2.42	0.00	960	204	135	1.51	0.00	69
7320725	CPNE2	270	111	2.43	0.00	159	94	54	1.74	0.00	40
5700128	MIAT	716	292	2.45	0.00	424	906	313	2.9	0.00	594
6480035	CTSA	1181	480	2.46	0.00	701	603	154	3.91	0.08	448
6350017	CNTNAP1	289	117	2.47	0.00	172	519	83	6.24	0.00	436
6520167	BIRC3	751	301	2.5	0.05	450	2038	768	2.65	0.07	1270
4610075	TFRC	567	221	2.57	0.00	346	1892	614	3.08	0.00	1278
2100328	YWHAH	776	300	2.59	0.00	476	2772	891	3.11	0.01	1881
2320301	ITGB1	2284	877	2.6	0.00	1407	4717	1853	2.55	0.00	2864
6480500	HLA-DPA1	1615	611	2.64	0.00	1004	1933	565	3.42	0.00	1368
7400136	HLA-DMA	669	253	2.64	0.00	416	1629	479	3.4	0.00	1150
7570440	E2F2	310	116	2.67	0.00	194	847	138	6.13	0.00	709
1440296	HLA-DQB1	292	109	2.69	0.00	183	157	50	3.11	0.27	106
4200037	PTPLA	317	117	2.71	0.00	200	315	53	5.94	0.00	262
3450685	LGALS3	272	100	2.71	0.02	172	508	80	6.38	0.01	428
2060148	TNFRSF4	790	290	2.72	0.00	500	93	64	1.46	0.02	30
770019	NINJ2	395	144	2.73	0.00	250	1163	242	4.81	0.00	921
990132	FUT7	367	132	2.77	0.00	235	1063	176	6.06	0.01	888
4670743	RNF214	1105	398	2.78	0.00	707	1099	245	4.49	0.00	855
6040600	NA	468	162	2.88	0.00	306	1052	168	6.26	0.00	884
1090064	CD74	1635	555	2.94	0.05	1080	3830	1088	3.52	0.08	2741
4540138	GBP5	1615	545	2.97	0.00	1070	2301	618	3.72	0.03	1683
3850451	STAM	407	137	2.98	0.00	270	618	109	5.7	0.00	510
6560066	BFSP2	359	116	3.09	0.00	243	175	56	3.13	0.00	119
5260750	FOXP3	294	93	3.14	0.00	200	73	64	1.14	0.02	9
20373	PTTG1	374	117	3.2	0.00	257	528	113	4.67	0.00	415
2710044	TRIB1	329	103	3.2	0.00	226	564	57	9.85	0.00	507
430338	ACTA2	520	155	3.36	0.00	365	672	113	5.96	0.00	559
2350324	PRDM1	668	197	3.39	0.01	471	1300	294	4.42	0.02	1006
2470471	DUSP4	424	123	3.45	0.00	301	457	68	6.72	0.00	389
1240491	SELP	417	116	3.6	0.00	301	229	61	3.74	0.00	168
5220538	IL10RA	1104	306	3.61	0.01	798	3732	788	4.74	0.03	2944
1400242	FANK1	370	102	3.64	0.00	268	264	50	5.26	0.00	214
4010053	HLA-DRB3	454	123	3.68	0.00	331	390	80	4.88	0.00	310
7100348	S100A4	6188	1635	3.79	0.00	4554	14602	4907	2.98	0.00	9696
3140561	S100A4	6188	1635	3.79	0.00	4554	18183	6061	3	0.00	12122
2710278	SHMT2	1386	363	3.81	0.00	1023	2935	608	4.83	0.01	2327
3780364	ANXA2	1112	289	3.84	0.00	823	104	58	1.78	0.02	45
1740164	PLEKHK1	469	121	3.87	0.00	348	310	52	5.94	0.00	258
7100348	S100A4	4160	1072	3.88	0.00	3087	14602	4907	2.98	0.00	9696
3140561	S100A4	4160	1072	3.88	0.00	3087	18183	6061	3	0.00	12122

2710577	NCF4	470	119	3.96	0.00	351	591	98	6.02	0.00	493
2750528	NA	808	203	3.98	0.00	605	979	135	7.24	0.00	844
1170307	IL2RB	1086	264	4.12	0.00	822	2869	506	5.67	0.00	2363
1230632	TNFRSF1B	1543	360	4.28	0.00	1183	4107	849	4.84	0.00	3258
5690382	METTL7A	429	99	4.32	0.00	330	119	43	2.79	0.00	77
2470110	JAKMIP1	664	153	4.33	0.00	510	55	53	1.03	0.39	2
6040379	HLA-DRB4	627	132	4.74	0.06	495	1355	132	10.24	0.00	1222
160377	LGALS1	1334	266	5.01	0.00	1068	3773	515	7.33	0.00	3258
2710577	NCF4	967	185	5.22	0.00	781	591	98	6.02	0.00	493
5090403	CTLA4	999	162	6.17	0.00	837	4397	372	11.82	0.00	4025

Supplementary Table 4 – Differentially expressed genes in the T_{reg} data set
Differentially expressed genes (FC > 1.75, p-value < 0.05, diff > 100) between T_{reg} and non-T_{reg} samples on I-huBC-V1 were calculated and corresponding values for these genes were checked on I-huBC-V2. Marked in grey are genes which show low signal values on I-huBC-V2.

Appendix C – Supplementary Methods

Stimulation of $CD4^+$ T cells

Cells were stimulated by mixing with artificial antigen-presenting cells (aAPCs) at a ratio of 1:3 (cells:beads) composed of magnetic beads (Dynal Biotech, Oslo, Norway) coated with the following antibodies: anti-CD3 (OKT3), anti-CD28 (9.3), anti–PD-1–17, and anti–MHC-I (W6/32). For all experiments, these aAPCs were coated with suboptimal anti-CD3Ab (5%), suboptimal levels of anti-CD28 Ab (14%), and either anti–MHC-I Ab (CD3/28/MHC-I) or anti–PD-1 Ab (CD3/28/PD-1), constituting the remaining 81% of protein added to the bead, as previously described.19 TGFβ was initially titrated at different concentrations ranging from 0 to 50 ng/mL to determine minimum concentration for maximum inhibitory effect in T-cell functions such as proliferation and cytokine production. For defining the TGFβ fingerprint, 30 ng/mL TGFβ was used. This concentration is also within the range of TGFβ described in serum derived from cancer patients of different origin (Shirai, Kawata et al. 1994; Toomey, Condron et al. 2001).

Cytometric bead array for cytokines

The concentration of IFN-γ in cell culture supernatants was measured using the human Th1/Th2 Cytokine kit II (BD Pharmingen, San Diego, CA) as described previously (Chemnitz, Driesen et al. 2006).

References

Abraham, R. T. and A. Weiss (2004). "Jurkat T cells and development of the T-cell receptor signalling paradigm." Nat Rev Immunol **4**(4): 301-8.

Alberts, B., D. Bray, et al. (2002). Molecular Biology of the Cell, Taylor & Francis.

Alizadeh, A. A., M. B. Eisen, et al. (2000). "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." Nature **403**(6769): 503-11.

Applied Biosystems "The Design and Annotation of the Applied Biosystems Human Genome Survey Microarray, White paper."

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Avery, O. T., C. M. MacLeod, et al. (1979). "Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III." J Exp Med **149**(2): 297-326.

Baechler, E. C., F. M. Batliwalla, et al. (2003). "Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus." Proc Natl Acad Sci U S A **100**(5): 2610-5.

Benson, D. A., I. Karsch-Mizrachi, et al. (2006). "GenBank." Nucleic Acids Res **34**(Database issue): D16-20.

Bentley, D. R. (2006). "Whole-genome re-sequencing." Curr Opin Genet Dev **16**(6): 545-52.

Bertucci, F., N. Borie, et al. (2004). "Identification and validation of an ERBB2 gene expression signature in breast cancers." Oncogene **23**(14): 2564-75.

Bhattacharjee, A., W. G. Richards, et al. (2001). "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses." Proc Natl Acad Sci U S A **98**(24): 13790-5.

Bild, A. H., G. Yao, et al. (2006). "Oncogenic pathway signatures in human cancers as a guide to targeted therapies." Nature **439**(7074): 353-7.

Bolstad, B. M., R. A. Irizarry, et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-93.

Bremnes, R. M., R. Sirera, et al. (2005). "Circulating tumour-derived DNA and RNA markers in blood: a tool for early detection, diagnostics, and follow-up?" Lung Cancer **49**(1): 1-12.

Canales, R. D., Y. Luo, et al. (2006). "Evaluation of DNA microarray results with quantitative gene expression platforms." Nat Biotechnol **24**(9): 1115-22.

Chemnitz, J. M., J. Driesen, et al. (2006). "Prostaglandin E2 impairs CD4+ T cell activation by inhibition of lck: implications in Hodgkin's lymphoma." Cancer Res **66**(2): 1114-22.

Chemnitz, J. M., D. Eggle, et al. (2007). "RNA fingerprints provide direct evidence for the inhibitory role of TGFbeta and PD-1 on CD4+ T cells in Hodgkin lymphoma." Blood **110**(9): 3226-33.

Chime Plugin. (2008). "http://www.mdl.com/products/framework/chime/."

Classen, S. (2008). TBA. Life and Medical Sciences Bonn, University of Bonn. **PhD Thesis**.

Classen, S., T. Zander, et al. (2007). "Human resting CD4+ T cells are constitutively inhibited by TGF beta under steady-state conditions." J Immunol **178**(11): 6931-40.

Crawley, M. J. (2007). The R Book. Chichester, Wiley & Sons.

Dahm, R. (2005). "Friedrich Miescher and the discovery of DNA." Dev Biol **278**(2): 274-88.

Dalgaard, P. (2001). The R-Tcl/Tk interface. 2nd International Workshop on Distributed Statistical Computing, Vienna, Austria.

DeRisi, J., L. Penland, et al. (1996). "Use of a cDNA microarray to analyse gene expression patterns in human cancer." Nat Genet **14**(4): 457-60.

Driesen, J. (2005). Einfluss von PGE2 auf die Aktivierung von primären CD4+ T Zellen. Internal Medicine I, University of Cologne. **Diploma Thesis**.

Dudoit, S., Y. H. Yang, et al. (2002). "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments." Statistica Sinica **12**: 111-139.

Dunning, M. J., M. L. Smith, et al. (2007). "beadarray: R classes and methods for Illumina bead-based data." Bioinformatics **23**(16): 2183-4.

Flicek, P., B. L. Aken, et al. (2007). "Ensembl 2008." Nucleic Acids Res.

Frueh, F. W. (2006). "Impact of microarray data quality on genomic data submissions to the FDA." Nat Biotechnol **24**(9): 1105-7.

Ganti, A. K. and J. L. Mulshine (2005). "Lung cancer screening: panacea or pipe dream?" Ann Oncol **16 Suppl 2**: ii215-9.

Garcia, A., X. Cayla, et al. (2003). "Serine/threonine protein phosphatases PP1 and PP2A are key players in apoptosis." Biochimie **85**(8): 721-6.

Gentleman, R. C., V. J. Carey, et al. (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biol **5**(10): R80.

Gentleman, R. C., V. J. Carey, et al. (2005). Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Berlin, Springer.

Golub, T. R., D. K. Slonim, et al. (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." Science **286**(5439): 531-7.

Gunderson, K. L., S. Kruglyak, et al. (2004). "Decoding randomly ordered DNA arrays." Genome Res **14**(5): 870-7.

Guo, L., E. K. Lobenhofer, et al. (2006). "Rat toxicogenomic study reveals analytical consistency across microarray platforms." Nat Biotechnol **24**(9): 1162-9.

Hardiman, G. (2006). "Microarrays Technologies 2006: an overview." Pharmacogenomics **7**(8): 1153-8.

Hoheisel, J. D. (2006). "Microarray technology: beyond transcript profiling and genotype analysis." Nat Rev Genet **7**(3): 200-10.

Huang, E., S. Ishida, et al. (2003). "Gene expression phenotypic models that predict the activity of oncogenic pathways." Nat Genet **34**(2): 226-30.

Huber, W., A. von Heydebreck, et al. (2002). "Variance stabilization applied to microarray data calibration and to the quantification of differential expression." Bioinformatics **18 Suppl 1**: S96-104.

Irizarry, R. A., D. Warren, et al. (2005). "Multiple-laboratory comparison of microarray platforms." Nat Methods **2**(5): 345-50.

Janeway, C. A., P. Travers, et al. (2005). Immunobiology, B&T.

Johnson, J. M., J. Castle, et al. (2003). "Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays." Science **302**(5653): 2141-4.

Kanehisa, M., M. Araki, et al. (2008). "KEGG for linking genomes to life and the environment." Nucleic Acids Res **36**(Database issue): D480-4.

Khan, J., R. Simon, et al. (1998). "Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays." Cancer Res **58**(22): 5009-13.

Khatri, P., P. Bhavsar, et al. (2004). "Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments." Nucleic Acids Res **32**(Web Server issue): W449-56.

Khatri, P., V. Desai, et al. (2006). "New Onto-Tools: Promoter-Express, nsSNPCounter and Onto-Translate." Nucleic Acids Res **34**(Web Server issue): W626-31.

Khatri, P. and S. Draghici (2005). "Ontological analysis of gene expression data: current tools, limitations, and open problems." Bioinformatics **21**(18): 3587-95.

Kim, J. B., G. J. Porreca, et al. (2007). "Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy." Science **316**(5830): 1481-4.

Kononen, J., L. Bubendorf, et al. (1998). "Tissue microarrays for high-throughput molecular profiling of tumor specimens." Nat Med **4**(7): 844-7.

Kronick, M. N. (2004). "Creation of the whole human genome microarray." Expert Rev Proteomics **1**(1): 19-28.

Kuhn, K., S. C. Baker, et al. (2004). "A novel, high-performance random array platform for quantitative gene expression profiling." Genome Res **14**(11): 2347-56.

Kuo, W. P., F. Liu, et al. (2006). "A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies." Nat Biotechnol **24**(7): 832-40.

Lander, E. S. (1996). "The new genomics: global views of biology." Science **274**(5287): 536-9.

Larkin, J. E., B. C. Frank, et al. (2005). "Independence and reproducibility across microarray platforms." Nat Methods **2**(5): 337-44.

Lesko, L. J. and J. Woodcock (2004). "Translation of pharmacogenomics and pharmacogenetics: a regulatory perspective." Nat Rev Drug Discov **3**(9): 763-9.

Lin, P., L. J. Medeiros, et al. (2004). "The activation profile of tumour-associated reactive T-cells differs in the nodular and diffuse patterns of lymphocyte predominant Hodgkin's disease." Histopathology **44**(6): 561-9.

Liu, W., A. L. Putnam, et al. (2006). "CD127 expression inversely correlates with FoxP3 and suppressive function of human CD4+ T reg cells." J Exp Med **203**(7): 1701-11.

Lockhart, D. J., H. Dong, et al. (1996). "Expression monitoring by hybridization to high-density oligonucleotide arrays." Nat Biotechnol **14**(13): 1675-80.

Mootha, V. K., C. M. Lindgren, et al. (2003). "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes." Nat Genet **34**(3): 267-73.

Mountain, C. F. (1997). "Revisions in the International System for Staging Lung Cancer." Chest **111**(6): 1710-7.

Mulshine, J. L. (2003). "Screening for lung cancer: in pursuit of pre-metastatic disease." Nat Rev Cancer **3**(1): 65-73.

Mumby, M. (2007). "PP2A: unveiling a reluctant tumor suppressor." Cell **130**(1): 21-4.

National Institute of Health. (2008). "What You Need To Know About Hodgkin lymphoma ".

Parkin, D. M., F. Bray, et al. (2005). "Global cancer statistics, 2002." CA Cancer J Clin **55**(2): 74-108.

Patterson, T. A., E. K. Lobenhofer, et al. (2006). "Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project." Nat Biotechnol **24**(9): 1140-50.

Patti, M. E., A. J. Butte, et al. (2003). "Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and NRF1." Proc Natl Acad Sci U S A **100**(14): 8466-71.

Pennacchio, L. A. and E. M. Rubin (2001). "Genomic strategies to identify mammalian regulatory sequences." Nat Rev Genet **2**(2): 100-9.

Petersen, K. F., S. Dufour, et al. (2004). "Impaired mitochondrial activity in the insulin-resistant offspring of patients with type 2 diabetes." N Engl J Med **350**(7): 664-71.

Pomeroy, S. L., P. Tamayo, et al. (2002). "Prediction of central nervous system embryonal tumour outcome based on gene expression." Nature **415**(6870): 436-42.

Pontius, J. U., L. Wagner, et al. (2003). UniGene: a unified view of the transcriptome. The NCBI Handbook. Bethesda (MD), National Center for Biotechnology Information.

Poppema, S. (1996). "Immunology of Hodgkin's disease." Baillieres Clin Haematol **9**(3): 447-57.

Poppema, S., M. Potters, et al. (1998). "Immune escape mechanisms in Hodgkin's disease." Ann Oncol **9 Suppl 5**: S21-4.

Pruitt, K. D., T. Tatusova, et al. (2007). "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic Acids Res **35**(Database issue): D61-5.

R Development Core Team (2007). R: A Language and Environment for Statistical Computing. Vienna, Austia, R Foundation for Statistical Computing.

R Development Core Team (2007). R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing.

Reinke, V. and K. P. White (2002). "Developmental genomic approaches in model organisms." Annu Rev Genomics Hum Genet **3**: 153-78.

Renzoni, E. A., D. J. Abraham, et al. (2004). "Gene expression profiling reveals novel TGFbeta targets in adult lung fibroblasts." Respir Res **5**: 24.

Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-70.

Schultze, J. L. and D. Eggle (2007). "IlluminaGUI: graphical user interface for analyzing gene expression data generated on the Illumina platform." Bioinformatics **23**(11): 1431-3.

Seddiki, N., B. Santner-Nanan, et al. (2006). "Expression of interleukin (IL)-2 and IL-7 receptors discriminates between human regulatory and activated T cells." J Exp Med **203**(7): 1693-700.

Shi, L., L. H. Reid, et al. (2006). "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." Nat Biotechnol **24**(9): 1151-61.

Shipp, M. A., K. N. Ross, et al. (2002). "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning." Nat Med **8**(1): 68-74.

Shippy, R., S. Fulmer-Smentek, et al. (2006). "Using RNA sample titrations to assess microarray platform performance and normalization techniques." Nat Biotechnol **24**(9): 1123-31.

Shirai, Y., S. Kawata, et al. (1994). "Plasma transforming growth factor-beta 1 in patients with hepatocellular carcinoma. Comparison with chronic liver diseases." Cancer **73**(9): 2275-9.

Siegel, P. M. and J. Massague (2003). "Cytostatic and apoptotic actions of TGF-beta in homeostasis and cancer." Nat Rev Cancer **3**(11): 807-21.

Smyth, G. K. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." Stat Appl Genet Mol Biol **3**: Article3.

Speed, T. P. (2003). Statistical Analysis of Gene Expression Microarray Data, Taylor & Francis Ltd.

Spellman, P. T., G. Sherlock, et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." Mol Biol Cell **9**(12): 3273-97.

Spira, A., J. E. Beane, et al. (2007). "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer." Nat Med **13**(3): 361-6.

Subramanian, A., P. Tamayo, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A **102**(43): 15545-50.

Swensen, S. J. (2003). "Screening for cancer with computed tomography." Bmj **326**(7395): 894-5.

Swensen, S. J., J. R. Jett, et al. (2005). "CT screening for lung cancer: five-year prospective experience." Radiology **235**(1): 259-65.

Tibshirani, R., T. Hastie, et al. (2002). "Diagnosis of multiple cancer types by shrunken centroids of gene expression." Proc Natl Acad Sci U S A **99**(10): 6567-72.

Tong, W., A. B. Lucas, et al. (2006). "Evaluation of external RNA controls for the assessment of microarray performance." Nat Biotechnol **24**(9): 1132-9.

Toomey, D., C. Condron, et al. (2001). "TGF-beta1 is elevated in breast cancer tissue and regulates nitric oxide production from a number of cellular sources during hypoxia re-oxygenation injury." Br J Biomed Sci **58**(3): 177-83.

Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A **98**(9): 5116-21.

Valk, P. J., R. G. Verhaak, et al. (2004). "Prognostically useful gene-expression profiles in acute myeloid leukemia." N Engl J Med **350**(16): 1617-28.

van 't Veer, L. J., H. Dai, et al. (2002). "Gene expression profiling predicts clinical outcome of breast cancer." Nature **415**(6871): 530-6.

van de Vijver, M. J., Y. D. He, et al. (2002). "A gene-expression signature as a predictor of survival in breast cancer." N Engl J Med **347**(25): 1999-2009.

Van Hoof, C. and J. Goris (2004). "PP2A fulfills its promises as tumor suppressor: which subunits are important?" Cancer Cell **5**(2): 105-6.

Vapnik, V. N. (1998). Statistical Learning Theory, Wiley-Interscience.

Velculescu, V. E. and K. W. Kinzler (2007). "Gene expression analysis goes digital." Nat Biotechnol **25**(8): 878-80.

Warnat, P., R. Eils, et al. (2005). "Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes." BMC Bioinformatics **6**: 265.

Watson, J. D. and F. H. Crick (1974). "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. J.D. Watson and F.H.C. Crick. Published in Nature, number 4356 April 25, 1953." Nature **248**(5451): 765.

Workman, C., L. J. Jensen, et al. (2002). "A new non-linear normalization method for reducing variability in DNA microarray experiments." Genome Biol **3**(9): research0048.

World Health Organization. (2008). "Cancer. World Health Organization."

Yagi, T., A. Morimoto, et al. (2003). "Identification of a gene expression signature associated with pediatric AML prognosis." Blood **102**(5): 1849-56.

List of Publications

Peer-reviewed publications

Chemnitz JM*, **Eggle D***, Driesen J, Classen S, Riley JL, Debey-Pascher S, Beyer M, Popov A, Zander T and Schultze JL

RNA fingerprints provide direct evidence for the inhibitory role of TGFβ and PD-1 on $CD4^+$ T cells in Hodgkin lymphoma, *Blood*. 2007 Nov 1;110(9):3226-33.

* shared first authorship

Classen S, Zander T, **Eggle D**, Chemnitz JM, Brors B, Büchmann I, Popov A, Beyer M, Eils R, Debey S and Schultze JL

Human resting $CD4^+$ T cells are constitutively inhibited by TGFβ under steady state conditions, *J Immunol.* 2007 Jun 1;178(11):6931-40.

Schultze JL and **Eggle D**

IlluminaGUI: Graphical User Interface for analyzing gene expression data generated on the Illumina platform, *Bioinformatics.* 2007 Jun 1;23(11):1431-3.

Popov A, Abdullah Z, Wickenhauser C, Saric T, Driesen J, Hanisch FG, Domann E, Raven EL, Dehus O, Hermann C, **Eggle D**, Debey S, Chakraborty T, Kronke M, Utermohlen O and Schultze JL

Indoleamine 2,3-dioxygenase-expressing dendritic cells form suppurative granulomas following Listeria monocytogenes infection, *J Clin Invest.* 2006 Dec;116(12):3160-70.

Abstracts at international meetings

Eggle D, Zander T, Brors B, Driesen J, Debey-Pascher S, Eils R and Schultze JL

GOAna: Assessing upstream events in gene expression microarray experiments using Gene Ontology, presented at *ISMB/ECCB 2007*, Vienna, Austria

Eggle D, Beyer M, Claßen S, Riley J, Debey-Pascher S and Schultze JL

First steps towards a systems biology approach of human regulatory T cells, presented at *World Immune Regulation Meeting 2007*, Davos, Switzerland

Debey-Pascher S, **Eggle D**, Zander T and Schultze JL

Use of transcriptional profiling for primary diagnosis as exemplified for AML: Where do we stand?, presented at the *NGFN Meeting 2006*, Heidelberg, Germany