Appendix

A Applications: Expression Analysis Projects

During my work I analyzed five different gene expression projects. In the first project I analyzed gene expression data of cardiomyopathy samples. The aim was to understand disease mechanisms better and derive a diagnostic microarray chip that can predict disease subtype. In the second project, I analyzed gene expression data of atheriosclerosis samples. The aim was to test the hypothesis that plaque smooth muscle cells have a distinct molecular phenotype from intima smooth muscle cells. The results of this project were published in Mulvihill et al. *(2004). In the third project I investigated the pathway of shiA in Shigella flexneri. It is known that shiA is directly related to the aggressiveness of infectivity. The aim was to find other genes involved in this pathway. In the fourth project, I analyzed gene expression data of marfan tissues. The aim was to derive a gene expression signature for marfan in cultured skin fibroblasts. The results of this study were submitted and are currently under review. In the fifth project, I analyzed the gene expression data of metastating and non metastating melanoma tissues. The goal was to find novel disease genes, identify novel subtypes, and derive a signature for the subtypes. The results of this study were submitted and are currently under review. Here, I will describe in more detail the cardiomyopathy project and the melanoma project. I will also point out the potential use of the data for the design of a diagnostic microarray chip.*

A.1 Cardiomyopathy project

This project was part of the Cardiovascular Disease Network Initiative for NGFN-2 (Nationales Genomforschungsnetz). Participating groups were the Department of Medicine, University of Heidelberg, Germany (Dr. Boris Ivandic, PD Dr. Dieter Weichenhan), the Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin (Prof. Dr. Patricia Ruiz, Thilo Storm), and the Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin (Dr. Rainer Spang and Jochen Jäger).

Cardiomyopathy is a medical disorder of the heart muscle leading to heart failure. The aim of this study was to distinguish between different cardiomyopathy etiologies, especially hypertrophic cardiomyopathy (HCM) and dilated cardiomyopathy (DCM). HCM results from an increased size of cardiac muscle cells (myocytes). Initially, the stroke volume is improved but soon neutralized by increased interstitial fibrosis and stiffness of the ventricles impairing relaxation during the diastolic phase. DCM results from an expansion of the ventricles leading initially to an improved pressure-volume relationship. However, over time, the ventricular blood volume increases and the heart muscle is overstrained and cannot pump efficiently any longer. This finally leads to heart failure. Both, hypertrophy and dilation, will eventually lead to irreversible changes of the heart physiology and cause additional myocardial strain. Due to the increased metabolic demands and impaired relaxation the blood supply to the heart muscle itself is impaired, causing further degeneration.

The 5-year survival rates of patients with chronic heart failure are comparable to those of cancer patients reflecting the limitations of our understanding of the causes and mechanisms of contractile dysfunction (Stewart *et al.*, 2001). Therefore, early diagnosis and accurate determination of the exact cardiac disease is essential for a successful treatment. So far, there have been several gene expression studies in primary cardiomyopathy, either hypertrophic or dilated, but none with the aim of developing a gene expression based microarray chip for routine pathological diagnosis (Friddle *et al.*, 2000; Ueno *et al.*, 2003; Kaynak *et al.*, 2003; Hwang *et al.*, 2002; Barrans *et al.*, 2002).

A.1.1 Project goals

In this project we aimed at the development of a custom microarray, called cardio chip, for the diagnostic assessment of gene expression in endomyocardial biopsies. So far, only genome wide gene expression approaches, including 8000-20000 mostly known genes, were conducted (Ueno *et al.*, 2003; Napoli *et al.*, 2003; Kaynak *et al.*, 2003; Barrans *et al.*, 2002). Yet, many clones are non-informative for the disease under study because they are not expressed at all in heart. For routine diagnosis this represents increased costs for microarray chip production and processing of the data. In a prestudy, we therefore screened a large non-redundant set of 76031 cDNA clones for informative clones with

respect to cardiac expression (Grzeskowiak *et al.*, 2003). From the 76031 cDNA clones, 11282 were found likely to be expressed in heart. A first custom research microarray was built using only these 11282 clones. With this microarray chip, called cardio chip 1 (CC1), 96 biopsies of early-stage heart disease patients were screened. Additionally, the expression profiles were analyzed with the aim to identify metabolic pathways and signal transduction to understand the translation of causes into mechanisms of remodeling and clinical phenotypes.

A.1.2 Methods

First, we describe how the custom array used for screening gene expression in cardiac tissues was designed. To select genes expressed in any cardiomyopathic or healthy individual, cDNA clones were identified as follows: Five pools of probes generated from amplified RNA (aRNA) from patients samples with different cardiomyopathies (9xDCM, 22xDCM, 8xHNOCM, 6xHOCM, 11xRSD) and human heart total RNA pooled samples were hybridized onto a comprehensive whole genome array. For this, we used the UniGene `RZPD2` (http://www.rzpd.de) set. This library was generated on the basis of the NCBI UniGene clustering and contained 76,031 cDNA clones, covering approximately 98% of the entire human genome. cDNA clones producing a ratio of signal to background intensities of >3 for total RNA and >7 for amplified RNA were chosen. A total of 11282 cDNA clones representing expressed genes in any of the hybridized samples were identified. This set of cDNA clones, designated CardioChip 1 (CC1), was then re-arrayed at the RZPD and spotted in duplicates onto Nylon membranes using in-house robots.

For clone based microarrays it was shown that almost a quarter of the clones spotted did not contain the sequence that was expected. Errors ranged from 20% (Ross *et al.*, 2000) to 21% (Poe *et al.*, 1991), to 23.5% (Kothapalli *et al.*, 2002). Therefore, for increasing our clone annotation quality all clones were subject to sequence verification. The bacterial clones of the CC1 cDNA set were grown overnight for isolation and sequencing. Sequencing templates were precipitated and sequenced using vector-specific primers and ABI3700 automated sequencing machines. In total 12609 clones were sequenced, some in replicates. The clones were filtered by a minimum phred value of 10 and short sequences of less than 70bp after vector clipping were discarded. In the end we obtained good quality sequences for 10676 of the 11282 cDNA clones. All these clones were clustered together with TIGR contigs and the resulting clusters were compared to UniGene (Version 092003), Swissprot (Version 092003), and Ensemble (Version 16.33). 7718 clones could hereby be characterized into 5732 clusters. Some of the clones could not be assigned to gene clusters because of different read length, 5' versus 3' sequencing, or contamination.

The CC1 cDNA arrays were subsequently used for analysis of gene expression in individual endomyocardial biopsies. Endomyocardial biopsies were routinely obtained to establish a diagnosis in patients with reduced left ventricular function in the absence of coronary artery disease (prior exclusion by selective coronary angiography) or in patients suffering

from vasculitis or rheumatoid arthritis, in whom myocardial affection had to be excluded. Biopsies were collected during heart catheterization after written informed consent was obtained in accordance with the medical and ethical guidelines of the Internal Review Board of the University of Lübeck. Endomyocardial samples (1-5 mm^3) were taken mostly from the septum wall of the right ventricle, snap frozen, and stored in liquid nitrogen until extraction of RNA. Human heart total RNA from Clontech and myocardial samples from non-transplanted donor hearts were used as controls.

In total we collected 69 biopsies of early-stage heart disease: 34 dilated cardiomyopathies (DCM), 15 hypertrophic cardiomyopathies (HCM), 6 heart transplants (HTX), one ischemic cardiomyopathy (ICM), two pulmonary myocarditis (PMK), 10 rheumatic systemic diseases (RSD) and one viral cardiomyopathy (VCM). Additionally 27 controls from healthy donor hearts were collected. From these 96 samples total RNA was isolated employing a standard Trizol (Invitrogen) method and subjected to two rounds of *in vitro* amplification using a modified version of the T7-based protocol (Eberwine (1996), MegaScript, Ambion). The concentration and integrity of the aRNA from each biopsy was measured using a 2100 Bioanalyzer (Agilent Technologies).

For hybridizations, 100 ng of aRNA or 1 μg of total RNA were used per cDNA array. Total RNA was poly dT-primed. The aRNAs were primed using random hexamers followed by reverse transcription in the presence of 33P-dCTP to produce radioactively labeled cDNA probes. Probes were purified using Sephadex columns, denatured and added to denatured salmon sperm and human placenta DNA as blocking reagents. This mixture was added to 10 ml hybridization buffer (1M NaCl, 1% SDS, 0.1xSSC) for overnight hybridization at 65°C. After three rounds of washing (0.1% SDS, 0.1X SSC) for 20 min at 65°C, nylon membranes were exposed overnight to Fuji BAS screens. Hybridization signal intensities were read as TIFF image files using a Fuji BAS 1800 reader. Image processing was done using AIDA, a software packages offering automatic spot detection after manual alignment of a grid to the spot images of the membranes. The resulting raw expression data was finally background corrected and normalized using VSN (Huber *et al.*, 2002).

A.1.3 Results

First, we used the twilight package (Scheid and Spang, 2005) to identify differential genes between the different disease entities. Between DCM and HCM, RSE and HCM, and RSE and DCM no differential genes were found at a FDR of 0.05 (Fig. A.1). However, between controls versus cardiomyopathies (DCM and HCM together), 2365 differential genes were identified. Between controls and DCM 2508 differential genes and between controls and HCM 929 differential genes were identified at a FDR of 0.05.

Additionally, we assessed the diagnostic microarray chip potential by running the *MCRestimate* procedure with linear SVMs that were trained in a balanced fashion. For each group we determined the best classification parameters in a complete cross validation scenario including gene selection using t-statistics. This provided us with the optimal number of

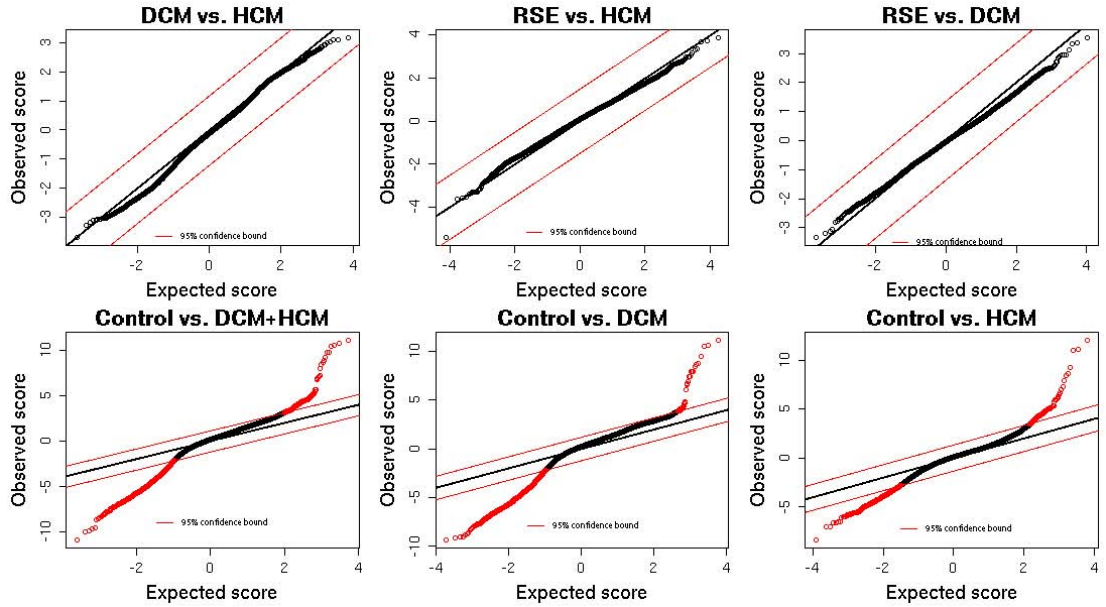

Figure A.1: Plot of observed t-statistics versus t-statistics of randomly permuted class labels. The dotted line show the 95% confidence interval calculated from permutations.

genes that were used in gene selection for the classifier in addition to the classification accuracy.

Predictive performances are generally contrasted with prevalence. The prevalence is the number of samples in the larger of the two group divided by the total number of samples. Thus, it is the fraction of correct classifications when assigning all samples to the bigger group.

For discriminating DCM from HCM the best classification accuracy was 55% (with a prevalence of 68%) using 10 genes. For discriminating DCM from controls the best classification accuracy was 97% (with a prevalence of 53%) using 20 genes. For discriminating HCM from controls the best classification accuracy was 98% (with a prevalence of 64%) using 5 genes. For discriminating both DCM and HCM from controls the best classification accuracy was 95% (with a prevalence of 64%) using 5 genes. Therefore, we conclude that a diagnostic microarray chip design for differentiating between the different cardiomyopathies is not possible with the probes screened.

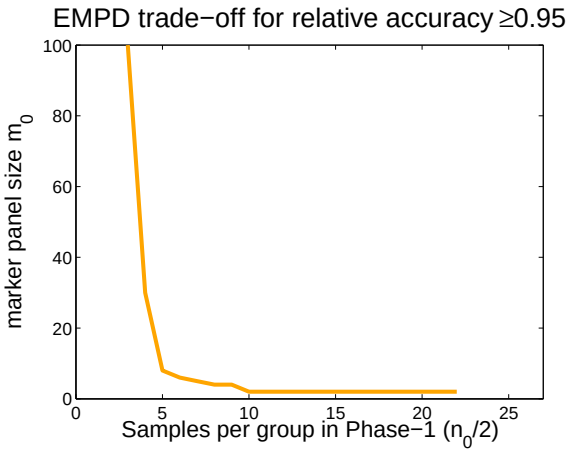

Figure A.2: Relationship between the number of genes in the marker panel and the number of samples examined in Phase-1 to achieve a relative accuracy of at least 95% ($A(n_0, m_0)/A(N-1, m_0) \geq 95\%$) in the Cardio dataset when distinguishing controls from cardiomyopathies.

However, it is possible to design a diagnostic microarray chip for discriminating between healthy patients and those suffering from an early-stage cardiomyopathy, which is relevant for routine clinical diagnosis. From our cross validation analysis we know that the optimal number of genes for classification is around 5-35. When using EMPD to assess how many patients should be screened to have no more than 5% loss in accuracy, we found that 5-10 patients per group are enough for deriving a marker panel (Fig. A.2).

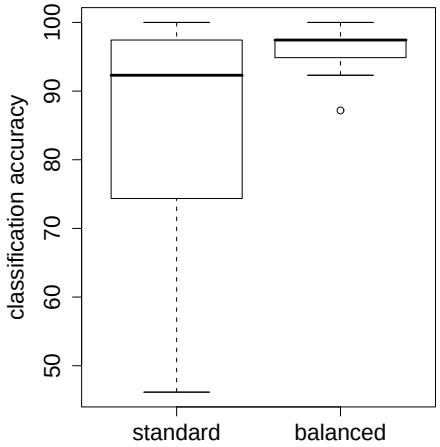

Figure A.3: Cross validation results of predictive performance of the same diagnostic signature used with balanced and standard normalization methods ($k_d = 5$, $k_n = 5$).

We finally assessed the normalization behavior of such a diagnostic microarray chip for cardiac diseases with 5 genes. Using standard normalization methods the accuracy dropped to 90 % but was recovered by using balanced normalization to 95% (Fig. A.3).

The number of differential genes between normal and cardiomyopathies is unexpectedly high. This can, however, reflect different covariates not related to etiology. Controls were heart transplants of healthy donors, while cardiomyopathy samples came from alive patients who already took medication. The potential of the diagnostic microarray chip has to be verified in a larger prospective or retrospective trial.

A.2 Melanoma project

This project was joint work from the Institute of Immunology, University of Rostock, Rostock, Germany; the Department of Dermatology and Venereology, University of Rostock, Rostock (Dr. Manfred Kunz); and the Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany (Dr. Rainer Spang and Jochen Jäger). It was funded by the Deutsche Krebshilfe, grant number 70-2819, and by NGFN (National Genome Research Network) grant number 031U209.

Melanoma is a skin cancer disease with a present lifetime risk of >1% in the caucasian population. The underlying factors for the development and progression of melanoma are largely unknown. Though, the central event in melanoma progression is distant metastasis. At this late stage of tumor progression only few therapeutic options exist with little impact on the patients overall prognosis. So far, the most important factor for the prognosis of the melanoma patient is vertical tumor thickness (Breslow, 1970). E.g. the average survival rate after ten years of a patient diagnosed with a 4mm thick melanoma is 40%. However, not only primary tumor thickness is of prognostic impact but in particular the transition from primary tumor to metastasis. Gene expression studies offer new insight into the underlying mechanisms and can lead to novel diagnostic tools. Clark *et al.* (2000) and Haqq *et al.* (2005) conducted gene expression profiles of metastatic and non-metastatic melanoma cell lines and identified progression factors arguing for a discriminative gene expression signature in primary tumors.

A.2.1 Project goal

In order to better understand the process of melanoma progression and metastasis, genome-wide gene expression profiling was performed. The goal was to find novel disease genes, identify novel subtypes, and investigate if clinical factors are correlated with gene expression patterns.

A.2.2 Methods

Tissue specimen of 19 primary melanomas and 22 cutaneous melanoma metastases were derived from 41 different patients. The biopsies were obtained after surgical excision of tumors. Diagnosis was made by routine histopathology. The vertical tumor thickness of primary melanomas ranged from 0.38 mm to 6.00 mm according to Breslow, including one in situ melanoma. Primary tumors consisted of different melanoma subtypes: superficial spreading melanomas (SSM, n=13), nodular melanomas (NM, n=5), and acrolentiginous melanomas (ALM, n=1). The metastatic lesions analyzed in this study were cutaneous or subcutaneous metastases. Biopsies were taken after tumors were surgically removed and immediately snap frozen in liquid nitrogen. Biopsies were preserved at -80°C until use for further analysis. The presented study was approved by the local ethic committee

of the University of Rostock and informed consent for microarray analyses of tumors was obtained from all patients.

We analyzed the tumor cell-specific gene expression of 22,283 probe sets using Affymetrix HG-U133a oligonucleotide microarrays. For this purpose tumor cells were excised from melanoma tissues by laser-capture microdissection in order to focus on melanoma cell-specific gene expression. The total RNA was extracted from tumor tissues using commercially available systems (RNeasy kit, Qiagen, Hilden, Germany). RNA concentrations were determined photometrically at 260 nm and RNA probes were labeled according to the suppliers instructions (Affymetrix, Santa Clara, CA, USA). 1 μg of total RNA was used for labeling. Hybridization and washing of gene chips were carried out according to the suppliers instructions. A laser scanner (Gene ArrayTM Scanner from Hewlett-Packard) was used for reading out the microarrays.

For data preprocessing, we performed background correction, normalization on probe level, and probe set summarization. The background correction was done similarly to Microarray Suite 5.0 (Affymetrix, 2001) but negative values were not truncated. Probe level normalization was done using the variance stabilization method by Huber *et al.* (2002). Finally, probe set summarization was performed using a median polish fit of an additive model (Irizarry *et al.*, 2003a).

In order to find genes with statistical significant differences in gene expression between different clinical phenotypes, genes were ranked according to the regularized t-score (Tusher *et al.*, 2001). False discovery rates (Storey, 2003) for the lists of top ranking genes were calculated based on 10,000 random permutations of the class labels. Differences in gene expression were regarded as significant when the false discovery rate (FDR) of the resulting lists did not exceed 0.05. Using the R Software for Statistical Computing (Ihaka and Gentleman, 1996) and the Bioconductor package GOstats every gene list was further examined for significant overrepresentation of biological processes and pathways defined by Gene Ontology (GO) categories. In order to derive diagnostic signatures, SVMs combined with a regularized t-score based feature selection filter were used. For compensating unbalanced group sizes we adjusted the class weights within the SVM according to the group sizes. Predictive performance was assessed using the *MCRestimate* package (Ruschhaupt *et al.*, 2004). Using *MCRestimate* the optimal number of genes and the optimal parameter setting of the SVM were determined in a nested cross validation setting. Cross validation was repeated 10 times with a 10-fold outer and 10-fold inner loop. We also implemented an in silico panel diagnosis that shows the patient-specific confidence of molecular diagnosis. This was achieved by running cross validation 1000 times with different random assignments to the 10-fold cross validation bins of patients and recording how often the SVM misclassified each sample. The SVM was trained using a linear kernel and feature selection of the best 100 genes.

A.2.3 Results

Overall 389 probe sets (representing 311 genes) were identified that showed significant differential expression between primary melanoma and melanoma metastases. A set of 57 probe sets (representing 47 genes) showed up-regulation in metastases and is of particular interest as tumor progression factor candidates. Significantly over-represented GO categories in the list of 311 genes were: cellular proliferation, cell cycle regulation, cell-cell contact, and cell-extracellular matrix interaction. The list contained several prominent genes including mitosin, CDC6, CDC28 (CDK1), septin 6, kinesin, osteopontin, and fibronectin. Further analysis showed that also subtypes of primary melanomas displayed characteristic gene expression patterns as did thin tumors (< 1.0 mm Breslow thickness) compared with thick tumors (> 2.0 mm Breslow thickness). Moreover, a predictive diagnostic model (support vector machine) for discriminating primary tumors and metastases was trained and a performance of more than 85% correct classifications was reached in cross validations. Currently, the results are submitted and under review.

It is still a matter of debate whether classification of primary melanomas into different sub-types, i.e. SSM, NM, ALM, or lentigo-maligna melanomas (LM), is more than a mere phenotypical classification. Thus, our next analysis addressed the question, whether melanoma sub-types may be differentiated by gene expression patterns. When comparing gene expression profiles of SSM and NM, a series of 67 probe sets (60 genes) was identified with a FDR < 0.05. This further supports the notion that there is indeed a clinical and molecular difference between both melanoma sub-types, and argues against the suggestion that the higher tumor thicknesses observed in NM may be the major factor that causes differences. The one ALM sample showed a gene pattern resembling that of metastases and NM which was verified by SVM classification (Fig. A.4). Further analyses showed that NM alone can not be differentiated from metastases based on gene expression patterns, which was expected based on the above mentioned findings indicating that three out of five NM showed gene patterns similar to that of metastases. In contrast, SSM can clearly be differentiated from metastases based on gene expression patterns. Functionally, SSM showed enhanced expression of genes involved in cell-cell contact and cell communication. The reduced expression of molecules involved in cell-cell contact observed in NM may be due to the fact that these tumors start vertical invasive growth, associated with loss of cell-cell contacts, immediately after malignant transformation.

The most important prognostic factor for malignant melanoma is the vertical tumor thickness. In order to address the question whether tumor thickness of primary tumors is represented by a particular gene expression pattern, thin primary tumors (< 1.0 mm Breslow thickness) were compared with thick primary tumors (> 2.0 mm Breslow thickness). When comparing both groups a list of 240 probe sets (representing 199 genes) was identified that showed differential expression with an estimated FDR of 0.05. Interestingly, 45 of the 60 differentially expressed genes (75%) in SSM versus NM overlapped with these 199 genes. Even more important, 116 of 199 (58%) of the list of differentially expressed genes overlapped with the list of differential genes of primary melanomas and metastases.

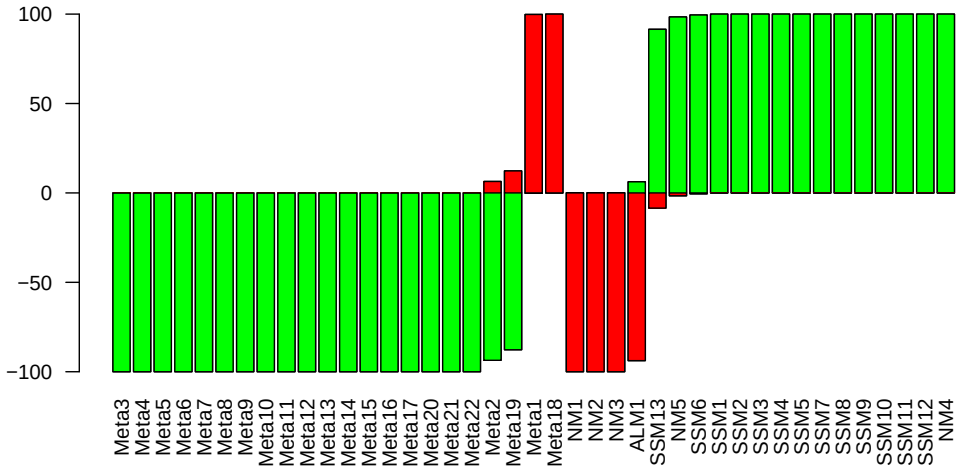

Figure A.4: Class prediction of primary melanomas and metastases using support vector machine analysis. Depicted are the percentage of misclassification for each tumor sample (primary melanoma or metastasis) when repeating 1.000 random 10-fold CV with a linear kernel SVM. The plot is sorted by misclassification rate and first shows all metastasis samples and then all melanoma samples. Light gray bars represent correct classification, dark Gray bars misclassification. Positive values indicate how often a sample was classified as melanoma and negative values how often it was classified as metastasis.

These findings argue for a progressive development of a metastatic signature in primary tumors coinciding with increasing tumor thickness. Moreover, tumors thicker than 2 mm showed gene expression patterns with a significant overlap to that of metastases, which supports the concept of tumor thickness as a major prognostic factor.

We next addressed the question whether gene expression profiles may accurately predict the tumor stage (primary melanoma or metastasis) of new unknown samples and calculated the estimated prediction accuracy. For this purpose, support vector machines (SVM) were applied. In an unbiased validation, using nested cross validation, our predictor reached an accuracy of 85% correct classifications (with a prevalence of 54 %). Classifiers of this accuracy can be designed with as little as 30 genes. In our in silico panel diagnosis (see Method section) mainly NM were misclassified and most of them were misclassified in all runs (Fig. A.4). The misclassified samples were NM1, NM2, NM3, Meta1 and Meta18. For a couple of patients (namely: Meta2, Meta19, ALM1 and SSM13) the SVM-diagnosis results were ambiguous between different cross validation runs, indicating that for these patients the expression profiles do not contain sufficient information to decide whether patients had primary tumors or metastases. The accuracy for distinguishing SSM and NM was 89% (with a prevalence of 72%) using 15 genes. The accuracy for distinguishing SSM and Meta was 91% (with a prevalence of 63%) using 50 genes. NM cannot be distinguished from Meta with an accuracy higher than prevalence.

These analyses showed that by use of a supervised classification method a high level of prediction accuracy for tumor stages may be achieved, which further argues for particular

biological stages represented by different melanoma subtypes or primary melanomas and metastases. SVM analyses using gene expression profiles can be used to differentiate between primary tumors and cutaneous metastases in the rare cases, when clinical and histopathological findings may not allow this differentiation.

In summary, based on gene expression profiles identified in our study, different sub-types of melanomas can be distinguished. This differentiation was further supported by the fact that gene expression profiles of SSM can be clearly differentiated from metastases, which was not possible for NM. Interestingly, two SSM with tumor thicknesses comparable to that of the NM showed gene pattern of SSM and not of NM. This argues against a differentiation between both sub-types merely based on different tumor thicknesses. These findings are in accordance with epidemiological data summarized in a recent publication on prognostic factors of melanoma, demonstrating that risk profiles of both tumor subtypes remain different even after correction for tumor thickness (Lomuto *et al.*, 2004). Using unsupervised and supervised methods for gene expression analysis in our study three NM showed gene patterns similar to metastases. Thus, NM may even be divided into different sub-types, one more closely related to metastases, and one more closely related to primary melanomas. At present, it cannot be decided, whether the three NM represent a different subtype of NM, or are on a further level of continuous de-differentiation. This question may be resolved by analyses of larger series of NM.

This study was also used to investigate the potential for designing a diagnostic microarray chip for differentiating metastating from non-metastating melanoma. First, we calculated the optimal number of genes needed for classification using *MCRestimate* as described above. For discriminating primary melanoma from melanoma metastases the best classification accuracy was 85% using 35 genes. Therefore we conclude that a diagnostic microarray chip for early detection of metastating melanomas is possible. This is relevant for clinical routine diagnosis and can improve melanoma therapy. When designing a microarray chip with 35 genes at least 6 samples should be screened in order to reach 95% relative accuracy of a whole genome assay (Fig. A.5).

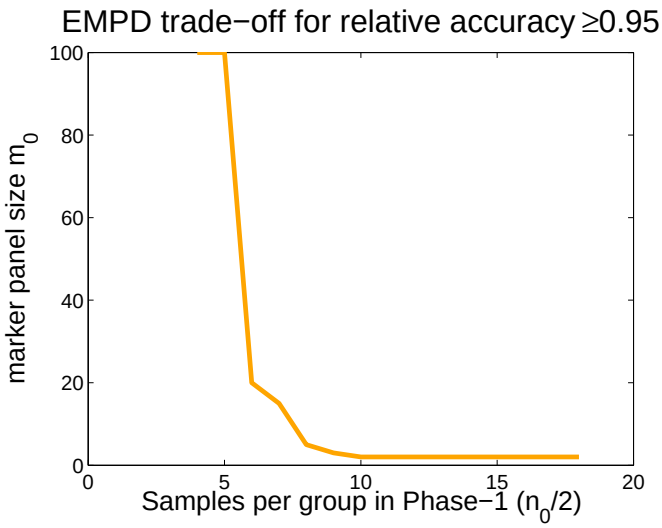

Figure A.5: Relationship between the number of genes in the marker panel and the number of samples examined in Phase-1 to achieve a relative accuracy of at least 95% ($A(n_0, m_0)/A(N-1, m_0) \geq 95\%$) when distinguishing metastating from non-metastating melanomas.

We finally assessed the normalization behavior of such a diagnostic microarray chip for

melanoma diseases with 15 genes. Using standard normalization methods the accuracy dropped to 77% but can be recovered by using balanced normalization to 80 % (Fig. A.6).

Taken together, the presented data demonstrate that melanoma metastasis represents a specific biological stage of tumor progression with a particular gene pattern. A majority of up-regulated genes in metastases fit the current pathogenic concepts of tumor progression and may serve as targets for innovative treatment approaches. Moreover, we were able to demonstrate that melanomas of different thicknesses and different melanoma sub-types are represented by particular gene expression patterns. Further studies should be initiated to analyze, whether gene expression patterns in primary melanomas may predict the prognosis of patients, and whether gene expression patterns in metastases may be used for treatment monitoring and designing a diagnostic microarray chip in clinical trials.

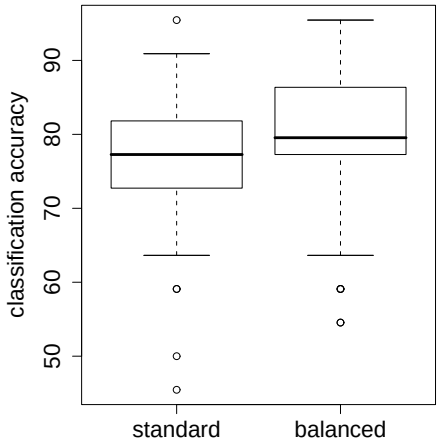

Figure A.6: Cross validation results of predictive performance of the same diagnostic signature used with balanced and standard normalization methods ($k_d = 15$, $k_n = 5$).