Statistical Applications in Genetics and Molecular Biology

Volume 4, *Issue* 1 2005 *Article* 9

Early Diagnostic Marker Panel Determination for Microarray Based Clinical Studies

Jochen Jaeger* Dieter Weichenhan†

Boris Ivandic‡ Rainer Spang**

*Max Planck Institute for Molecular Genetics, Berlin, jochen.jaeger@molgen.mpg.de

†Universitätsklinikum Heidelberg, Internal Medicine III, dieter.weichenhan@med.uni-heidelberg.de

‡Universitätsklinikum Heidelberg, Internal Medicine III, Boris.Ivandic@med.uni-heidelberg.de

**Max Planck Institute for Molecular Genetics, Berlin, rainer.spang@molgen.mpg.de

Early Diagnostic Marker Panel Determination for Microarray Based Clinical Studies*

Jochen Jaeger, Dieter Weichenhan, Boris Ivandic, and Rainer Spang

Abstract

We present a novel, cost efficient two-phase design for predictive clinical gene expression studies: early marker panel determination (EMPD). In Phase-1, genome-wide microarrays are used only for a small number of individual patient samples. From this Phase-1 data a panel of marker genes is derived. In Phase-2, the expression values of these marker panel genes are measured for a large group of patients and a predictive classification model is learned from this data. Phase-2 does not require the use of expensive whole genome microarrays, thus making EMPD a cost efficient alternative for current trials. The expected performance loss of EMPD is compared to designs which use genome-wide microarrays for all patients. We also examine the trade-off between the number of patients included in Phase-1 and the number of marker genes required in Phase-2. By analysis of five published datasets we find that in Phase-1 already 16 patients per group are sufficient to determine a suitable marker panel of 10 genes, and that this early decision compromises the final performance only marginally.

KEYWORDS: medical diagnostics, diagnostic chip design, microarray, gene expression, marker genes

---

*We would like to thank Claudio Lottaz and Florian Markowetz for careful proof reading of the manuscript and stimulating discussions. This work was supported by NGFN (National Genome Research Network) grant 01GS0108.

1 Introduction

Recent publications demonstrated the high potential of gene expression studies using microarrays for the diagnosis of tumor entities (Bhattacharjee *et al.*, 2001; Yeoh *et al.*, 2002), the determination of risk groups (Huang *et al.*, 2003; van't Veer *et al.*, 2002), and the prediction of the response to treatment (Cheok *et al.*, 2003). The challenge to associate expression patterns with clinical disease phenotypes is still active focus of current research. In microarray based clinical studies, tissue samples are characterized by high dimensional vectors, containing the expression levels of thousands of genes. These datasets typically consist of tens or hundreds of samples (patients), but contain several thousand variables (expression levels of genes). This setting, with more variables than samples, leads to saturated models. Additional constraints or regularization techniques are required to derive predictive classification models that generalize well. A widely used regularization approach is variable selection, i.e., constraining classification models to only a few, selected variables. In the context of microarray classification, those variables are marker genes. Several methods for variable selection were proposed by Golub *et al.* (1999); Ben-Dor *et al.* (2000); Dudoit *et al.* (2002); Roth *et al.* (2002) and compared in Dudoit *et al.* (2002); Jaeger *et al.* (2003). On the one hand, variable selection serves the statistical purpose of deriving classification models that generalize well (Tibshirani *et al.*, 2002). On the other hand, it allows for a cost efficient study design as we will show in this paper. Our main finding is that the sample-size requirements for variable selection are lower than those for classifier learning, i.e., few samples can be used efficiently to determine a set of variables on which a classifier is further fine-tuned using a large number of samples.

Current clinical studies collect whole genome data of all patients screened. Then they apply gene selection to the complete dataset. In this paper we suggest a novel two step approach to which we refer to as Early Marker Panel Determination (EMPD). In the first step (Phase-1), genome-wide microarrays are used to screen a small number of patients only and to derive a diagnostic marker panel from this data. In the second step (Phase-2), the expression values of these marker genes only are measured in a large group of patients. This dataset is used for calibrating the final predictive model. Thus EMPD is less expensive because expression analysis of a small set of genes can be done very cost efficiently using alternative quantification methods like quantitative reverse transcription PCR (qRT-PCR, Heid *et al.* (1996)). However, since less data is available for variable selection we will loose predictive performance.

The paper is organized as follows: In section 2 we describe the subsampling based evaluation procedure to determine the expected performance loss caused by EMPD. In section 3 we evaluate the loss of performance by EMPD analyzing five publicly available datasets. We conclude with a summary of our findings and discuss their implications for the design of clinical microarray studies.

2 A subsampling approach to evaluate the effect of EMPD

Since published datasets for the two phase design of EMPD are not available, we exploit data from large clinical whole genome studies. We simulate Phase-1 by randomly choosing a

subset of n_0 patients for which we use the complete expression profiles determined on whole genome microarrays. From this data we determine the marker panel. To simulate Phase-2 we ignore all non-marker-panel genes. With the expression values of marker panel genes obtained from Phase-1 and Phase-2, we finally determine a classification model. All datasets in this paper can be divided into two groups of patients. More formally, let N be the total number of samples in a dataset, with $N/2$ samples in each group. Let M be the total number of genes on the microarray used during Phase-1. After having analyzed a subset of $n_0 < N$ patients with $n_0/2$ samples in each group, we decide on a small set of genes $m_0 \ll M$. To account for sample variance effects, we randomly draw 30 sample subsets $S_i, i \in \{1, .., 30\}$ of size n_0 without replacement. Analyzing only the patients in S_i we derive a virtual marker panel P_i containing m_0 genes. Finally, we train a multivariate classification model using the complete set of samples but analyzing only genes from the panel P_i. We evaluate the performance of this classifier denoting the prediction accuracies by $A_i(n_0, m_0) = (N - E_i)/N$, where E_i is the number of misclassifications. In total, this gives us 30 accuracy values $A_i(n_0, m_0)$ for each combination of values n_0 and m_0. We denote $A(n_0, m_0)$ as the median of these 30 values. To estimate the performance of EMPD, we compare $A(n_0, m_0)$ to the leave-one-out estimate $A(N-1, m_0)$, which reflects the performance of the traditional approach including all patients in the analysis. Note, that we cannot unbiasedly compare to N samples. At least one sample has to be left out as a test set.

The evaluation of classifier performance is nontrivial. Several papers have pointed out possible pitfalls leading to over optimistic estimators (Ambroise and McLachlan, 2002; West *et al.*, 2001; Chatfield, 1995). To avoid the feature selection bias described in Ambroise and McLachlan (2002), we use external leave-one-out cross validation (LOOCV) where in each step feature selection is performed separately. Iteratively, we set aside each sample as a test sample, then we randomly draw $n_0/2$ samples for each group from the remaining samples. On these n_0 samples we determine the m_0 marker panel genes. Using these genes only on $N - 1$ samples, we train a Support Vector Machine (SVM). This SVM then classifies the left-out sample. After each sample has been left out in turn we obtain N classification results and compare them to the known labels to determine the error rates E_i (Fig. 1).

On randomized class labels, this procedure gives the expected prevalence of 50% (data not shown). To estimate variability, Mukherjee *et al.* (2003) pointed out that the observed variance of classifier performances is higher than the expected population variance but the quantiles of the leave-one-out estimator are unbiased. We therefore use boxplots showing quantiles in figures 2 and 3. For simplicity, we only apply two standard variable selection procedures. The marker panels P_i consist of the m_0 genes with the highest two-sample t-statistic or Wilcoxon rank sum statistic, respectively, in S_i. The related problem of how to select markers has been addressed in the literature before. For a comparative study of several feature selection methods and classifiers see Lee *et al.* (2005) and Jaeger *et al.* (2003). Subsequent model fitting is done using SVMs with radial basis function kernels (Gist 1.3β with default parameters; `http://microarray.genomecenter.columbia.edu/gist/`). To evaluate EMPD for 10 different choices of m_0 and n_0, respectively, on one dataset with 128 samples, the procedure needs 24 hours CPU time parallelized on 8 Athlon 1.8GHz machines.

Main Function:
foreach m_0 = number of markers
foreach n_0 = number of samples in Phase-1
for $i \in \{1, .., 30\}$ repeats
calculate $A_i(n_0, m_0)$

Subroutine:
$A_i(n_0, m_0) \leftarrow$ **function(...)**
let $E = 0$ # Errors made so far
foreach sample $d \in D = \{1, .., N\}$ # LOOCV
put d as test sample aside
$S \leftarrow$ draw n_0 samples from $D \backslash \{d\}$ in a balanced fashion
$P \leftarrow$ determine marker panel as top m_0 markers of S
train SVM with $D \backslash \{d\}$ samples on P markers
test d, restricted to marker panel P, with learned SVM classifier
if classification is wrong then increment E
return $(N - E)/N$

Fig. 1. Pseudo code for EMPD evaluation procedure

3 Applications of EMPD

We examine five published datasets (Tab. 1). All five datasets use Affymetrix HGU95Av2 DNA chips containing 12625 probesets, corresponding to more than 9000 known, unique, human genes. For preprocessing, we perform background correction, normalization on probe level, and probeset summarization. The background correction is done similarly to MAS 5 (Affymetrix, 2001) but negative values are not truncated. Probe level normalization is done using the variance stabilization method by Huber *et al.* (2002). Finally, probeset summarization is performed using a median polish fit of an additive model described in Irizarry *et al.* (2003). For simplicity, we focus on classification problems with only two possible outcomes and randomly omit samples to obtain balanced sample numbers in each group.

The first dataset is a study on acute lymphocytic leukemia (ALL) in children (Yeoh *et al.*, 2002). 327 leukemia samples fall into different clinical classes characterized by immunophenotype, chromosomal translocations and aberrations. In this paper we focus on the diagnosis of hyper-diploid B-cell leukemias, a moderately complicated diagnostic problem. Using a balanced subset of all 64 samples displaying hyper-diploidy with more than 50 chromosomes and 64 samples randomly chosen from the rest of the samples, we achieve a LOOCV performance of 96% correct SVM classifications. The second dataset consists of 102 tumor and normal prostate tissues (Singh *et al.*, 2002). We obtain 92% accuracy for the classification of 50 tumor versus 50 normal tissues. Furthermore, we examine a dataset of lung cancer samples (Bhattacharjee *et al.*, 2001), where 98% accuracy for the classification of 21 squamous carcinomas versus 21 adenocarcinomas is achieved. The last dataset contributed by Huang *et al.* (2003) consists of 89 breast cancer samples which are divided into a study for recurrence (34 non-recurrent and 18 recurrent patients, further denoted as breastR) and a study for lymph-node risk (18 high-risk and 19 low risk samples, further denoted as breastL). In this prognosis setting we classify 92% of the samples correctly using

SVM on 18 recurrent versus 18 samples randomly chosen from the non-recurrent pool. In the lymph-node risk study, 65% of the samples were classified correctly. The later is a hard classification task, achieving a performance slightly above random guessing.

We describe the results of the first dataset in detail and only summarize corresponding results of the four other datasets in tables 2 and 3. To evaluate EMPD, we first examine

Dataset	Group 1: sample size	Group 2: sample size
Leukemia (Yeoh *et al.*, 2002)	Hyper-diploid: 64	Other B-cells: 64 of 200
Prostate (Singh *et al.*, 2002)	Normal: 50	Tumor: 50 of 52
Lung (Bhattacharjee *et al.*, 2001)	Squamous: 21	Adenocarcinomas:21 of 190
BreastR (Huang *et al.*, 2003)	Recurrent: 18	Non-recurrent: 18 of 34
BreastL (Huang *et al.*, 2003)	High risk: 18	Low Risk: 18 of 19

Table 1
Datasets used for the evaluation of EMPD. Groups 1 and 2 denote the groups used for the evaluation with EMPD and their sample sizes.

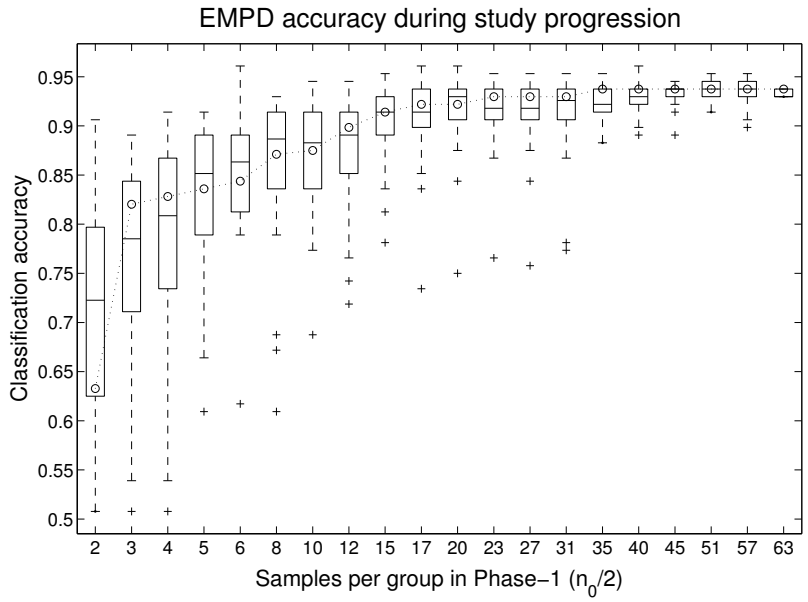

Fig. 2. Accuracy of EMPD for a marker panel of 10 genes applied to the leukemia (Yeoh *et al.*, 2002) dataset. The boxplots refer to analysis using t-statistic and show the distribution of classification accuracies ($A_i(n_0, 10), i = \{1, .., 30\}$) for 30 subsamplings. The dotted line and the circles refer to the Wilcoxon statistic and show median accuracies only. The x-axis is in polynomial scale.

the loss of prediction accuracy for a fixed marker panel size. For a marker panel of 10 genes, less than 20 samples in Phase-1 are sufficient to reach saturating performances (Fig. 2). Such a small marker panel may readily be examined by qRT-PCR. Without EMPD, we observe a median accuracy of $A(N-1, 10) = 93\%$. As expected, EMPD reduces the median accuracy and increases its variance. However, except for extremely small sample sizes in Phase-1, the loss in accuracy appears to be marginal. Even with only 12 patients per group we get $A(12 * 2, 10) = 89\%$ corresponding to 96% relative accuracy i.e., accuracy in relation to standard classification that uses all data for the feature selection (relative

accuracy $= A(n_0, m_0)/A(N-1, m_0)$). The use of relative accuracies allows a comparison of the EMPD results of datasets with different final classification power. Note, that these results do not differ notably when using a Wilcoxon or t-statistic.

There is a trade-off between the number of patients used in Phase-1 and the size of the marker panel. Larger marker panels can achieve state of the art performance with only a few patients in Phase-1. For a fixed Phase-1 sample size of $n_0 = 10 * 2$ and varying marker panel sizes m_0, satisfying results cannot be achieved with panel sizes of 2 and 3 markers (Fig. 3). However, already 10 genes lead to an absolute accuracy of 88%. With 30 genes the absolute accuracy reaches 92%. Using more genes ($m_0 = 100$) increases the accuracy to 94% which corresponds to a relative accuracy of 99%.

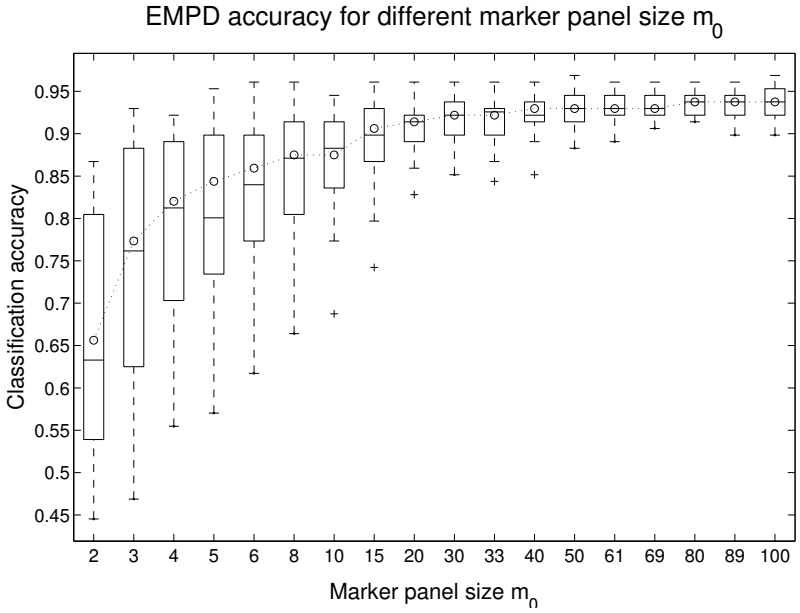

Fig. 3. Accuracy of EMPD for the leukemia dataset (Yeoh *et al.*, 2002) when different marker panel sizes m_0 are evaluated. The number of samples in Phase-1 is fixed to 10 patients in each group ($n_0 = 10 * 2$). The boxplots refer to analysis using t-statistic and show the distribution of SVM leave-one-out cross validation accuracies across 30 runs of random patient subsampling. The dotted line and the circles refer to the Wilcoxon statistic and show median accuracies only.

The results suggest that there is a direct sample size - panel size trade-off. Using more marker genes facilitates a Phase-1 with less samples, whereas more samples in Phase-1 permit a smaller marker panel. We have determined the number of genes required to reach a relative accuracy of $\geq 95\%$ for a Phase-1 with a given number of n_0 samples (Fig. 4). The corresponding plots for the other 4 studies are similar (data not shown). EMPD can therefore be used to determine the number of necessary marker genes for a given Phase-1 size. Vice versa, it can be used to determine the number of samples needed in Phase-1 for a given marker panel size.

We determined the relative accuracy of EMPD with a marker panel size of $m_0 = 10$ genes and $m_0 = 100$ genes. For the small marker panel with $m_0 = 10$ genes, a very small Phase-1 with 5 patients is enough to achieve $\geq 92\%$ relative performance for the lung and the leukemia dataset. When doubling Phase-1 to 10 patients per group, already four datasets achieve $\geq 95\%$ relative accuracy. Only the breastR dataset needs more samples in Phase-1

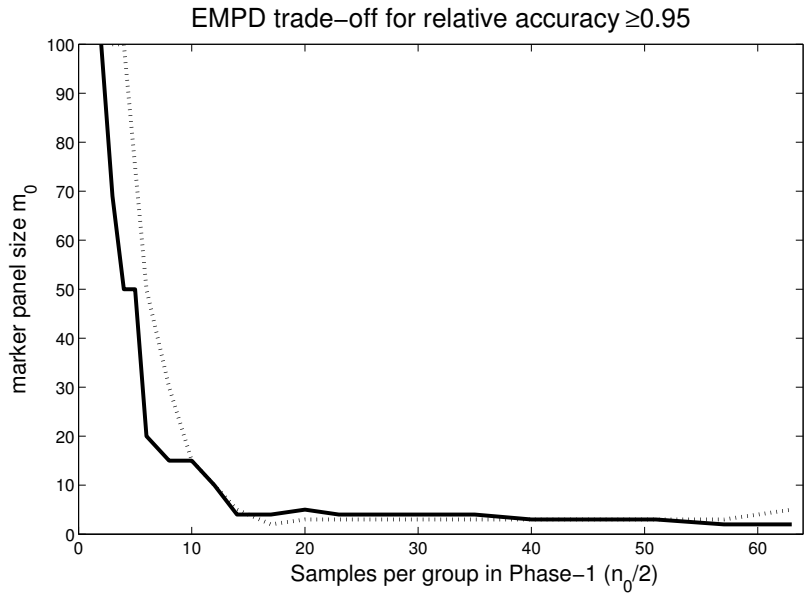

Fig. 4. Relationship between the number of genes in the marker panel and the number of samples examined in Phase-1 to achieve a relative accuracy of at least 95% ($A(n_0, m_0)/A(N-1, m_0) \geq 95\%$) in the leukemia dataset (Yeoh *et al.*, 2002). The dotted line depicts the curve when using a Wilcoxon test statistic, the solid line when using a two sample t-statistic.

and achieves $\geq 94\%$ relative accuracy with 15 patients in Phase-1 (Table 2). When using $m_0 = 100$ all datasets but the lung dataset achieve relative accuracies $\geq 99\%$ with only 10 patients per group. The advantage of EMPD is that it can successfully accommodate both, a limited number of genes in the marker panel as well as a limited number of samples to be screened in Phase-1.

	Panel with $m_0 = 10$			Panel with $m_0 = 100$		
$n_0/2$	5	10	15	5	10	15
leukemia	92%	95%	98%	98%	99%	100%
prostate	76%	95%	97%	93%	99%	100%
lung	95%	99%	100%	93%	95%	98%
breastL	85%	100%	100%	100%	100%	100%
breastR	73%	86%	94%	90%	100%	100%

Table 2
Relative classification accuracy of EMPD ($A(n_0, m_0)/A(N-1, m_0)$). Accuracies are calculated using SVM leave-one-out cross-validation.

We found that the leukemia and the lung data allow good predictive performance with an extremely small Phase-1 even for a marker panel with just 10 genes. For the leukemia dataset, 12 patients, for the breastL dataset 9 patients and for the lung dataset 3 patients are sufficient to achieve $\geq 95\%$ relative accuracy. For the prostate and breastR cancer dataset, a larger Phase-1 (15 and 16 patients) is needed (Table 3).

We also investigated the overlap of marker panels across 30 runs of random subsampling. The good performance of EMPD suggests that there are many informative genes and that prediction can be based on very different combinations of them. Especially for small n_0 the

	Samples needed for 95% relative performance		
Dataset	N/2	Panel with $m_0 = 10$	Panel with $m_0 = 100$
leukemia	64	12	2
prostate	50	15	6
lung	21	3	2
breastL	18	9	4
breastR	18	16	10

Table 3
Comparison of sample requirements for EMPD, with a small ($m_0 = 10$) or a medium size ($m_0 = 100$) marker panel, to achieve at least 95% relative accuracy. Accuracies are calculated by standard SVM leave-one-out cross-validation. $N/2$ denotes the total number of samples per group in the datasets.

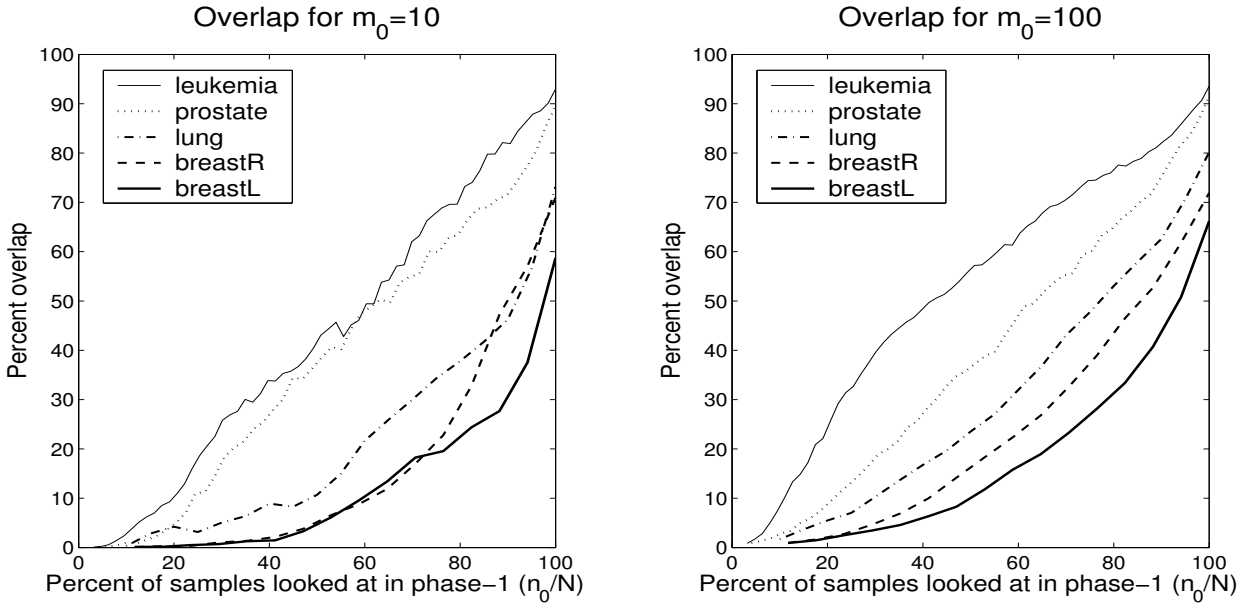

Fig. 5. Mean pairwise overlap of the marker panels in the 30 subsamplings for each n_0 and $m_0 = \{10, 100\}$

marker panels hardly overlap at all. For harder classification tasks the overlap is in general smaller and increases slower in n_0 (Fig. 5).

4 Discussion

We propose a novel, two step study design for clinical gene expression profiling studies. For a small number of patients whole genome microarray data is collected (Phase-1). Then a marker panel is determined from this Phase-1 data. From now on, this marker panel is used to screen a large patient pool (Phase-2). Furthermore, we introduce a novel evaluation procedure to determine the loss in classification accuracy depending on the number of patients in Phase-1 and the size of the marker panel.

Analyzing five published clinical microarray datasets we find that in Phase-1 as little as 16 patients per group are sufficient to identify a panel of 10 marker genes. For a marker

panel of 100 genes, not more than 10 patients per group are needed. The early decision on the marker panel compromises the final performance of the diagnostic classification only marginally. We show that there is an inverse relationship between the number of samples in Phase-1 and the size of the marker panel. Using more samples in Phase-1 facilitates the identification of a more reliable set of markers. Therefore, fewer markers are sufficient to achieve the same relative performance. On the other hand, if it is possible to use many markers, only few samples need to be screened in Phase-1.

Our results demonstrate that EMPD is a feasible design for cost efficient clinical studies based on gene expression levels. Material, production and handling costs can be saved. Since only few genes in Phase-2 need to be examined, it is possible to utilize small custom diagnostic mRNA arrays or other technologies like qRT-PCR, in-situ hybridization or protein panels (Büssow *et al.*, 2001). These technologies may also be closer to the clinical phenotype (protein panel) or more precise (qRT-PCR).

It is important to note that we obtain different marker panels using different subsets of patients for EMPD without a noticeable loss of classification accuracy. Notably, a small sample size of only 10-20 patients may not be enough to determine the most comprehensive set of discriminating genes. However, for a good classification performance it is not necessary to identify those genes. It is not even necessary that all genes in the panel are informative marker genes. In many cases, a few informative genes in the panel are enough to obtain a strong signature at the end of Phase-2. Our observation that finding marker genes for classification is easy and does not require many patients suggests that there are many, probably up to thousands of informative genes in all five studies. In fact, estimating the number of differentially expressed genes using the method by Scheid and Spang (2004) indicates several thousand differentially expressed genes in all four studies, too. However, it is unclear whether the molecular cause of the clinical phenotypes involves several thousand genes. On the other hand, classification does not need to identify causes. Genes involved in secondary and tertiary effects are valuable molecular markers as well. While these marker genes may serve well for diagnostic purposes, they may not be useful to elucidate the molecular basis of a disease and many of them can be replaced by equally well performing marker genes.

While our results show that the relative accuracy after EMPD is only slightly compromised even for problems with a poor overall performance, it is clear that EMPD can not improve absolute performance. If the absolute performance without EMPD is insufficient for practical use, EMPD is of no use too.

This paper is purely descriptive, and our findings only apply to the five datasets shown. However, since our results in all five studies are consistent, we believe that EMPD is appropriate for other clinical studies as well and even ongoing studies may benefit. But for a study that has no fixed sample size target N, it is not clear how to determine the optimal length of Phase-1 or the optimal number of marker genes. From this perspective, it would be helpful to have a computational tool to guide EMPD during a running genome-wide study. Our evaluation procedure does not enable us to do this. In an ongoing study, performance at the end of Phase-2 cannot be evaluated, but needs to be extrapolated from the available Phase-1 data. Mukherjee *et al.* (2003) introduced a method for sample size estimation using powerlaw extrapolation. This approach can be extended to EMPD as well and needs to be further investigated.

References

Affymetrix (2001) *Microarray Suite User Guide, Version 5.0.*

Ambroise,C. and McLachlan,G. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A,* **99** (10), 6562–6.

Ben-Dor,A., Bruhn,L., Friedman,N., Nachman,I., Schummer,M. and Yakhini,Z. (2000) Tissue classification with gene expression profiles. *J Comput Biol,* **7** (3-4), 559–83.

Bhattacharjee,A., Richards,W., Staunton,J., Li,C., Monti,S., Vasa,P., Ladd,C., Beheshti,J., Bueno,R., Gillette,M., Loda,M., Weber,G., Mark,E., Lander,E., Wong,W., Johnson,B., Golub,T., Sugarbaker,D. and Meyerson,M. (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.

Büssow,K., Konthur,Z., Lueking,A., Lehrach,H. and Walter,G. (2001) Protein array technology. Potential use in medical diagnostics. *Am J Pharmacogenomics,* **1** (1), 37–43.

Chatfield,C. (1995) Model uncertainty, data mining and statistical inference. *J. R. Stat. Soc. Series A,* **158**, 419–66.

Cheok,M., Yang,W., Pui,C., Downing,J., Cheng,C., Naeve,C., Relling,M. and Evans,W. (2003) Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. *Nat Genet,* **34** (1), 85–90.

Dudoit,S., Yang,Y.H., Callow,M.J. and Speed,T.P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica,* **12**, 111–39.

Golub,T., Slonim,D., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J., Coller,H., Loh,M., Downing,J., Caligiuri,M., Bloomfield,C. and Lander,E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science,* **286** (5439), 531–7.

Heid,C., Stevens,J., Livak,K. and Williams,P. (1996) Real time quantitative PCR. *Genome Res,* **6** (10), 986–94.

Huang,E., Cheng,S., Dressman,H., Pittman,J., Tsou,M., Horng,C., Bild,A., Iversen,E., Liao,M., Chen,C., West,M., Nevins,J. and Huang,A. (2003) Gene expression predictors of breast cancer outcomes. *Lancet,* **361** (9369), 1590–6.

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

Irizarry,R., Hobbs,B., Collin,F., Beazer-Barclay,Y., Antonellis,K., Scherf,U. and Speed,T. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics,* **4** (2), 249–64.

Jaeger,J., Sengupta,R. and Ruzzo,W. (2003) Improved gene selection for classification of microarrays. In *Pacific Symposium on Biocomputing 2003* pp. 53–64 Pac Symp Biocomput World Scientific.

Lee,J., Lee,J., Park,M. and Song,S. (2005) An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis,* **48**, 869–85.

Mukherjee,S., Tamayo,P., Rogers,S., Rifkin,R., Engle,A., Campbell,C., Golub,T. and Mesirov,J. (2003) Estimating dataset size requirements for classifying DNA microarray data. *J Comput Biol,* **10** (2), 119–42.

Roth,V., Lange,T., Braun,M. and Buhmann,J. (2002) A resampling approach to cluster validation. In *Proc of COMPSTAT* Proc of COMPSTAT Physica Verlag.

Scheid,S. and Spang,R. (2004) A stochastic downhill search algorithm for estimating the local false discovery rate. *IEEE Transactions on Computational Biology and Bioinformatics,* **1** (3), 98–108.

Singh,D., Febbo,P., Ross,K., Jackson,D., Manola,J., Ladd,C., Tamayo,P., Renshaw,A., D'Amico,A., Richie,J., Lander,E., Loda,M., Kantoff,P., Golub,T. and Sellers,W. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell,* **1** (2), 203–9.

Tibshirani,R., Hastie,T., Narasimhan,B. and Chu,G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A,* **99** (10), 6567–72.

van't Veer,L., Dai,H., van de Vijver,M., He,Y., Hart,A., Mao,M., Peterse,H., van der Kooy,K., Marton,M., Witteveen,A., Schreiber,G., Kerkhoven,R., Roberts,C., Linsley,P., Bernards,R. and Friend,S. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature,* **415** (6871), 530–6.

West,M., Blanchette,C., Dressman,H., Huang,E., Ishida,S., Spang,R., Zuzan,H., Olson,J., Marks,J. and Nevins,J. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A,* **98** (20), 11462–7.

Yeoh,E., Ross,M., Shurtleff,S., Williams,W., Patel,D., Mahfouz,R., Behm,F., Raimondi,S., Relling,M., Patel,A., Cheng,C., Campana,D., Wilkins,D., Zhou,X., Li,J., Liu,H., Pui,C., Evans,W., Naeve,C., Wong,L. and Downing,J. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell,* **1** (2), 133–43.