2 Early marker panel determination (EMPD)

On the one hand, feature 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 microarray study design as I will show in this chapter. In this chapter, I propose a novel setup for microarray based clinical trials that exploits feature selection. First, I use gene expression data from whole genome studies for deriving a small set of diagnostically relevant genes (Phase-1). Then, a diagnostic microarray holding these genes is designed. Next, additional samples are screened with this diagnostic microarray (Phase-2). With this phase-2 data a classifier aiming for diagnosis is fine tuned. Furthermore, I assess the accuracy of classification with diagnostic microarrays depending on the number of samples used for determining the diagnostic biomarkers, on the number of biomarkers used, and on the number of samples provided for fine tuning the diagnostic microarray. A biomarker is a gene that is used in a diagnostic signature. Screening more samples in phase-1 before deriving the diagnostic microarray improves classification performance in general. However, the loss in performance quickly converges. The results show that it is possible to switch to a small diagnostic microarray after screening only a few samples in phase-1 without sacrificing substantial classification accuracy.*

In general, the classification accuracy improves as the number of training samples increases. On the other hand, measuring more samples means also increased costs for the production and evaluation of the microarrays needed. This cost increase is almost linear but the improvement in classification accuracy levels down quickly following an inverse power law (Mukherjee *et al.*, 2003). Hence, there is a natural point when increasing the training sample size does not improve accuracy significantly any more. In this thesis we go one step further and analyze how the number of samples in phase-1 and the number of probes measured by the microarray influence the classification accuracy.

Using miniaturized diagnostic microarray instead of whole genome microarrays helps to lower costs. A whole microarray holds substantially more probes, which means increased material costs. One benefit of switching from a whole genome microarray to a miniaturized diagnostic microarray are substantial expenditure savings. An Affymetrix GeneChip Human Genome U133 Plus 2.0 analyzing 47400 transcripts costs e.g. 975 US$, whereas a custom express array from the same company costs 375 US$ (Affymetrix Price Sheet 2006, Retail). On the other hand, the hybridization itself is also more expensive because larger microarrays also require more reagents. Furthermore, small custom microarrays

also allow a faster evaluation. First, there is less data to store in databases again reducing disc costs. Second, the number of genes to derive a diagnostic signature is much smaller speeding up the classification algorithms.

Therefore, we do not only consider how many samples should be screened but in parallel how many features should be put on a diagnostic microarray chip in order to have an acceptable classification accuracy. We suggest a novel two step approach 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. The data derived from this larger group is used for calibrating the final predictive model. Thus, EMPD is very effective because expression analysis of a small set of genes can be done cost efficiently using alternative methods like qRT-PCR. However, since less data is available for feature selection we will loose predictive performance.

Several papers have been published for assessing sample size requirements for microarray experiments (Pan *et al.*, 2002; Müller *et al.*, 2004; Tsai *et al.*, 2005; Zien *et al.*, 2003). However, most sample size calculations treat genes independently and do not aim at classification but at identifying a small number of truly differentially expressed genes. For developing a classifier from gene expression data it is more appropriate to measure the accuracy of the whole classifier on an independent test set or in cross validation. Methods for assessing the accuracy of classification depending on the sample size have been proposed (Mukherjee *et al.*, 2003). However, we put this one step further by considering additionally the switch from whole genome microarrays to small diagnostic microarray during such a study (Jäger *et al.*, 2005).

The chapter is organized as follows: In section 2.1, we describe the subsampling based evaluation procedure to determine the expected performance loss caused by EMPD. In section 2.2, we evaluate the loss of performance with EMPD by analyzing five publicly available datasets. In section 2.3, we investigate the relation between sample size in phase-1 and the number of markers used on the diagnostic microarray chip. Finally, in section 2.4, we show the overlap in the selected marker genes depending on the number of samples analyzed in phase-1. We conclude with a summary of our findings and discuss their implications for the design of clinical microarray studies.

2.1 A subsampling approach to evaluate the effect of EMPD

We exploited data from five large clinical whole genome studies to mimic our proposed two phase design of EMPD. We simulated phase-1 by randomly choosing a subset of n_0 patients for which we used the complete expression profiles determined on whole genome microarrays. From this data we determined the marker panel. To simulate phase-2 we ignored all non-marker-panel genes. With the expression values of marker panel genes

obtained from phase-1 and phase-2, we finally determined a classification model. More formally, let N be the total number of samples in a dataset, with $N/2$ samples in each group. Let P 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, a small set of relevant genes $p_0 \ll P$ is selected. To account for sample variance effects, we drew r subsets $S_i, i \in \{1, .., r\}$ of size n_0. Each S_i was randomly sampled without replacement from all cases. For our experiments $r = 30$ was chosen. Analyzing only the patients in S_i we derived a virtual marker panel M_i containing p_0 genes from all samples. Finally, we trained a multivariate classification model using the complete set of samples but analyzing only genes from the panel M_i. We evaluated the performance of the classifier denoting the prediction accuracies by $A_i(n_0, p_0) = (N - E_i(n_0, p_0))/N$, where $E_i(n_0, p_0)$ is the number of misclassifications. In total, this gave us 30 classification accuracy values $A_i(n_0, p_0)$ for each combination of values n_0 and p_0. We denote $A(n_0, p_0)$ as the median of these 30 values. To estimate the performance of EMPD, we compared $A(n_0, p_0)$ to the leave-one-out estimate $A(N - 1, p_0)$, which reflects the performance of an approach including all patients in the analysis. Note, that it is not possible to evaluate the performance using all N samples in the training set. At least one sample has to be left out as a test set.

2.1.1 Classifier evaluation

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 used external leave-one-out cross validation (LOOCV) where in each cross validation fold feature selection was performed separately. Iteratively, we set aside each sample as a test sample, then we randomly drew $n_0/2$ samples for each group from the remaining samples. On the n_0 samples we determined the p_0 marker panel genes. Using these genes only on $N - 1$ samples, we trained a Support Vector Machine (SVM). With the SVM we then classified the left-out sample. After each sample was left out in turn we obtained N classification results and compared them to the known labels to determine the error rates E_i (Fig. 2.1).

To estimate classification performance variability, Mukherjee *et al.* (2003) pointed out that the observed variance of classifier performances is higher than the expected population variance and therefore optimistic. Using quantiles of the leave-one-out estimator give an accurate estimate of a classifier trained with all but one sample and tested on an independent sample. We therefore used boxplots showing quantiles in our following figures. For simplicity, we only applied two standard feature selection procedures. The marker panels M_i consisted of the p_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 introductory chapter 1.4.3. Subsequent model fitting was done using SVMs with radial basis function kernels. We used the SVM implementation `Gist` (http://microarray.genomecenter.columbia.edu/gist) 1.3β with default parameters. To

Main Function:
foreach p_0 = number of markers
foreach n_0 = number of samples in phase-1
for $i \in \{1, .., r\}$ repeats
calculate $A_i(n_0, p_0)$

Subroutine:
$A_i(n_0, p_0) \leftarrow$ **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
$M \leftarrow$ determine marker panel as top p_0 markers of S
train SVM with $D \backslash \{d\}$ samples on the marker panel M
test d, restricted to marker panel M, with learned SVM classifier
if classification is wrong then increment E
return $(N - E)/N$

Figure 2.1: Pseudo code for EMPD evaluation procedure

evaluate EMPD for 10 different choices of p_0 and n_0, respectively, on one dataset with 128 samples, the procedure took 24 hours CPU time parallelized on 8 Athlon 1.8GHz machines.

Applications of EMPD

We examined five published datasets (Tab. 2.1). All five datasets used Affymetrix HGU95Av2 DNA chips containing 12625 probesets, corresponding to more than 9000 known, unique, human genes. For preprocessing, we performed background correction, normalization on probe level, and probeset summarization. The background correction was done similarly to MAS 5 (Affymetrix, 2001) but negative values were not truncated. Probe level normalization was done using the variance stabilization method by Huber *et al.* (2002). Finally, probeset summarization was performed using a median polish fit of an additive model described in Irizarry *et al.* (2003b). For simplicity, we focused on classification problems with only two possible outcomes and randomly omitted samples to obtain balanced sample numbers in each group.

The first dataset analyzed was 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 the analysis we focused 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 achieved a LOOCV performance of 96%. The second dataset consisted of 102 tumor and normal prostate tissues (Singh *et al.*, 2002). We obtained 92% accuracy for the classification of 50 tumor versus 50 normal tissues. Furthermore, we examined a dataset of lung cancer samples (Bhattacharjee *et al.*, 2001), where 98% accuracy for the classification of 21 squamous carcinomas versus 21 adenocarcinomas was achieved. The last dataset contributed by Huang *et al.* (2003) consisted 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 classified 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 was a hard classification task, achieving a performance slightly above random guessing.

2.2 EMPD results for four gene expression studies

We describe the results of the first dataset in detail and only summarize corresponding results of the four other datasets.

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

First, we examined 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 were sufficient to reach saturating performances (Fig. 2.2). Without EMPD, we observed a median accuracy of $A(N-1,10) = 93\%$ for the leukemia dataset. As expected, EMPD reduced the median accuracy and increases its variance. However, except for extremely small sample sizes in phase-1, the loss in accuracy appeared to be marginal. In the prostate and the lung dataset saturation happened very soon, whereas in the breastL and especially breastR dataset it is not clear if saturation was already reached (Fig. 2.3).

Next, we evaluate the performance of EMPD by comparing relative accuracies, which are defined as the accuracy in relation to classification that used all data for the feature selection: relative accuracy $= A(n_0, p_0)/A(N-1, p_0)$. The use of relative accuracies allows a comparison of the EMPD results of datasets with different final classification accuracy. Even with only 12 patients per group in the leukemia dataset we got $A(12*2, 10) = 89\%$

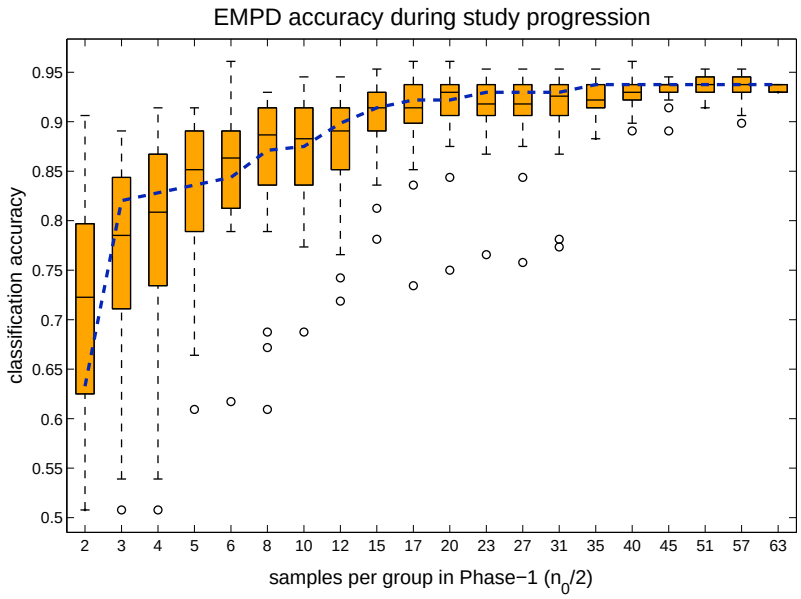

Figure 2.2: Accuracy of EMPD for a marker panel of 10 genes applied to the leukemia (Yeoh *et al.*, 2002) dataset. The boxplots refer to analysis using t-statistic and show the distribution of classification accuracies $(A_i(n_0, 10), i = \{1, .., 30\})$ for 30 subsamplings. The dotted line refers to the Wilcoxon statistic and shows median accuracies only. The x-axis is in polynomial scale.

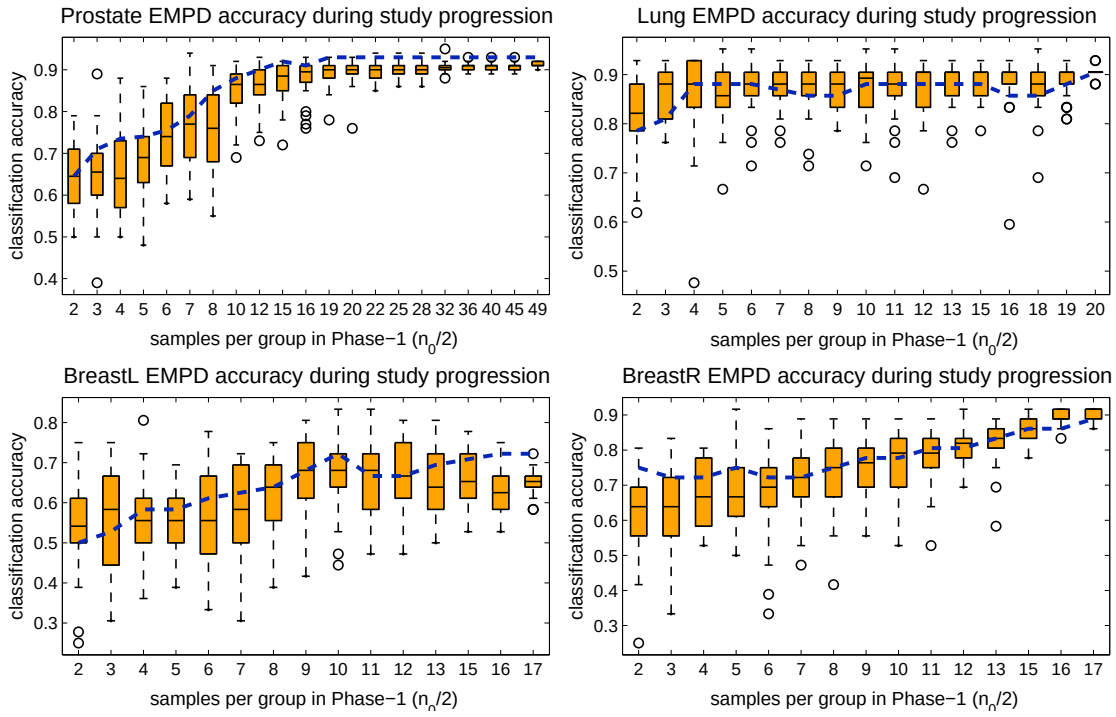

Figure 2.3: Accuracy of EMPD for a marker panel of 10 genes applied to the prostate, lung, breastR and breastL dataset. Legend see Fig. 2.2

corresponding to 96% relative accuracy. Note, that the classification results did not differ notably when using a Wilcoxon or t-statistic.

2.3 Relation between sample size and number of screened genes

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 p_0, satisfying results could not be achieved with panel sizes of 2 and 3 markers (Fig. 2.4). However, already 10 genes lead to a relative accuracy of 93%. With 30 genes the relative accuracy reached 97%. Using more genes ($p_0 = 100$) increased the accuracy to a relative accuracy of 99%.

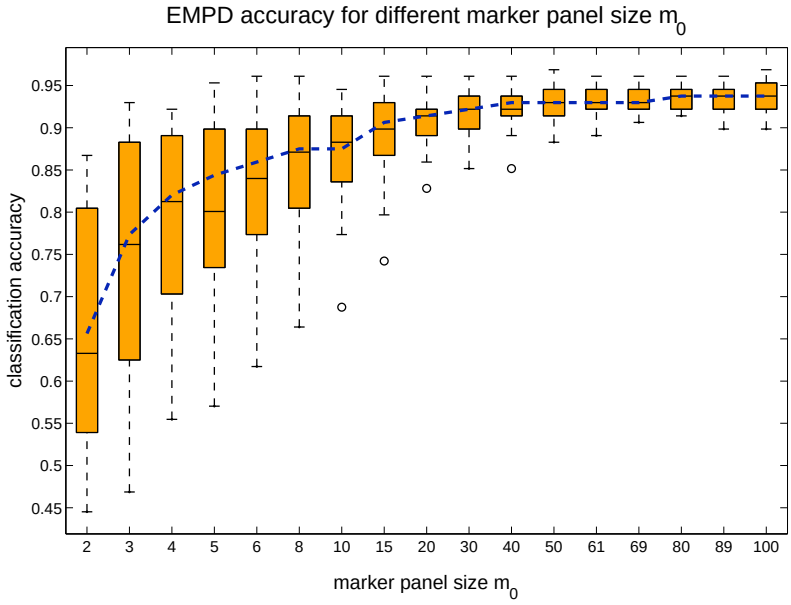

Figure 2.4: Accuracy of EMPD for the leukemia dataset (Yeoh *et al.*, 2002) when different marker panel sizes p_0 are evaluated. The number of samples in phase-1 is fixed to 10 patients in each group ($n_0 = 10 * 2$). The boxplots refer to analysis using t-statistic and show the distribution of SVM leave-one-out cross validation accuracies across 30 runs of random patient subsampling. The dotted line refers to the Wilcoxon statistic and shows 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. 2.6). The corresponding plots for the lung, prostate, breastL and breastR study are shown in Fig. 2.3, Fig. 2.5, and Fig. 2.7.

EMPD can therefore be used to determine the number of necessary marker genes for a

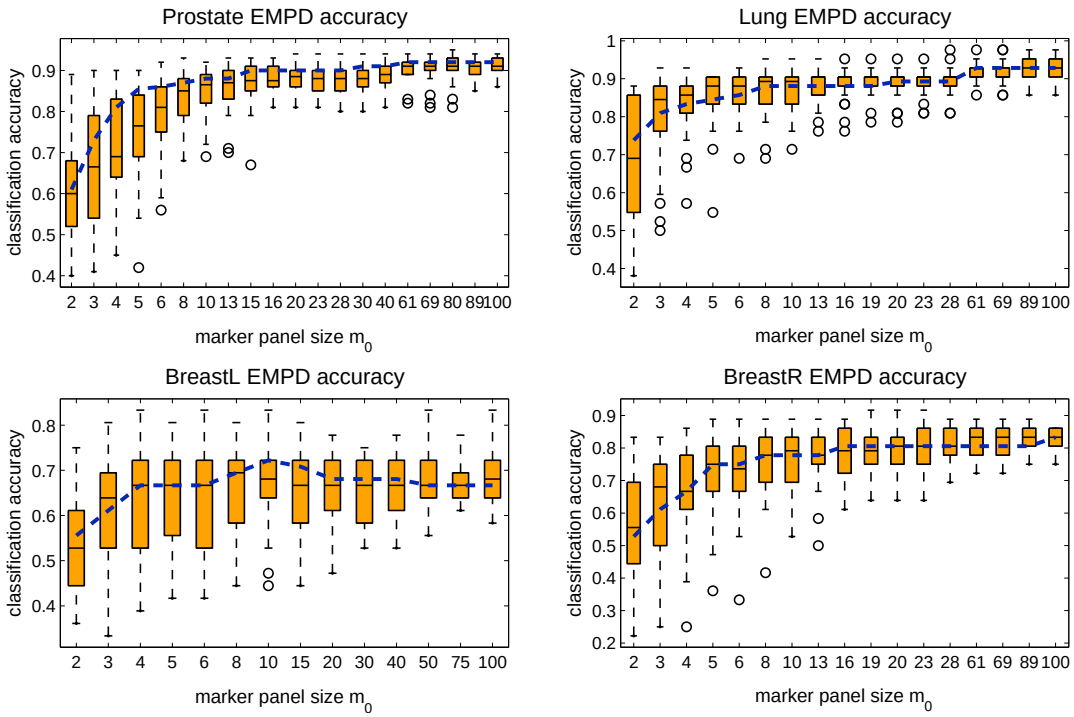

Figure 2.5: Accuracy of EMPD for the the prostate, lung, breastL and breastR data when different marker panel sizes p_0 are evaluated. Legend and further description see Fig. 2.4.

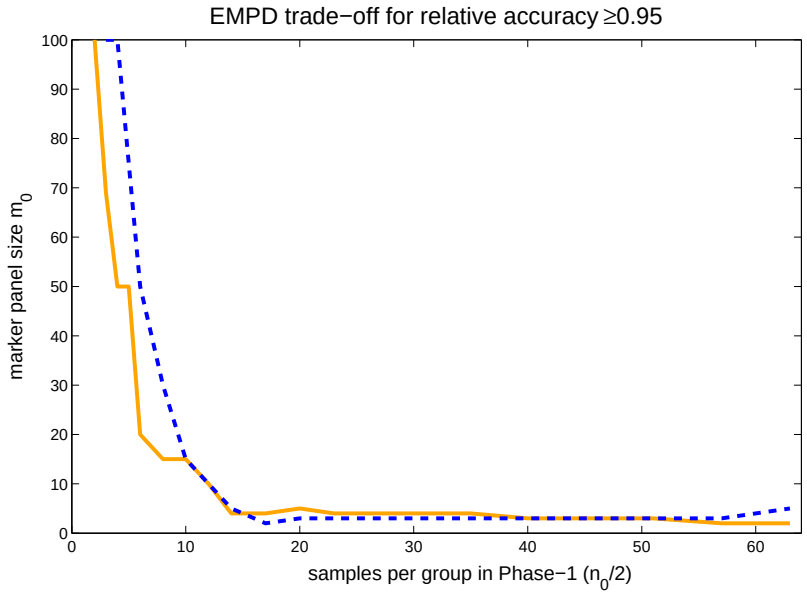

Figure 2.6: Relationship between the number of genes in the marker panel and the number of samples examined in phase-1 to achieve a relative accuracy of at least 95% ($A(n_0, p_0)/A(N-1, p_0) \geq 95\%$) in the leukemia dataset (Yeoh *et al.*, 2002). The dotted line depicts the curve when using a Wilcoxon test statistic, the solid line when using a two sample t-statistic.

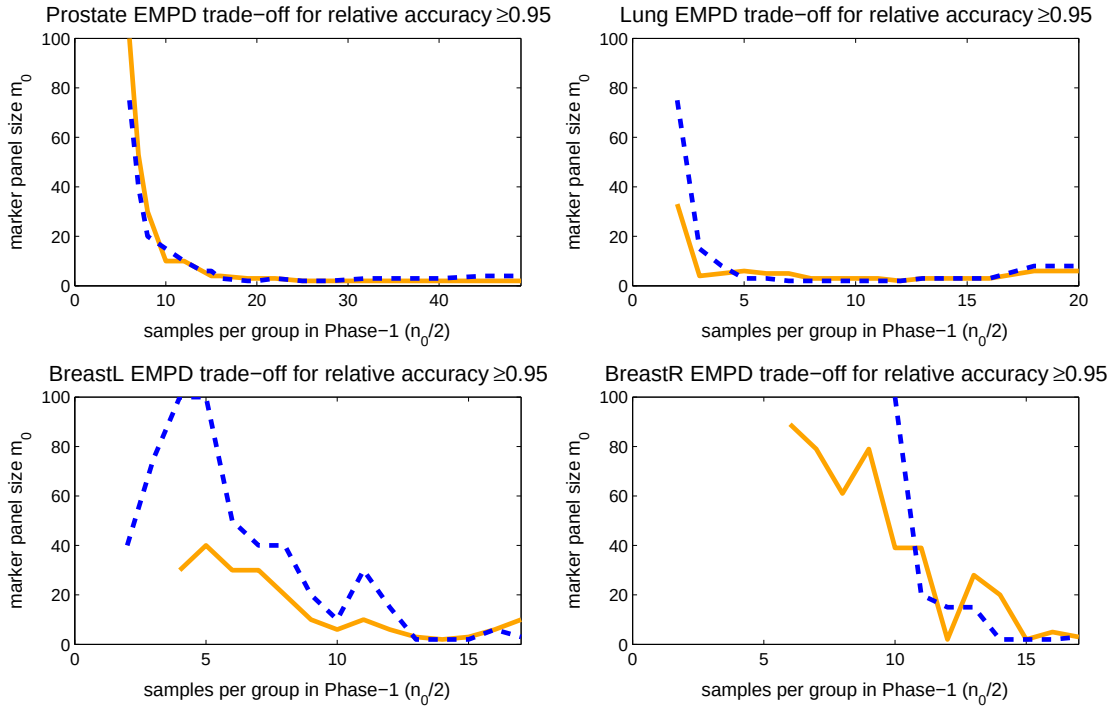

Figure 2.7: Relationship between the number of genes in the marker panel and the number of samples examined in phase-1 to achieve a relative accuracy of at least 95% ($A(n_0, p_0)/A(N-1, p_0) \geq 95\%$) for the the prostate, lung, breastL and breastR data. Legend see Fig. 2.6.

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 $p_0 = 10$ genes and $p_0 = 100$ genes. For the small marker panel with $p_0 = 10$ genes, a small phase-1 with 5 patients was 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 achieved $\geq 95\%$ relative accuracy. Only the breastR dataset needed more samples in phase-1 and achieves $\geq 94\%$ relative accuracy with 15 patients in phase-1 (Table 2.2). When using $p_0 = 100$ all datasets but the lung dataset achieved 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.

We found that the leukemia and the lung data allowed good predictive performance with a very 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 were sufficient to achieve $\geq 95\%$ relative accuracy. For the prostate and breastR cancer dataset, a larger phase-1 (15 and 16 patients) was needed (Table 2.3).

	Panel with $p_0 = 10$			Panel with $p_0 = 100$		
$n_0/2$	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 2.2: Relative classification accuracy of EMPD ($A(n_0, p_0)/A(N-1, p_0)$). Accuracies were calculated using SVM leave-one-out cross-validation.

	Samples needed for 95% relative performance		
Dataset	N/2	Panel with $p_0 = 10$	Panel with $p_0 = 100$
leukemia	64	12	2
prostate	50	15	6
lung	21	3	2
breastL	18	9	4
breastR	18	16	10

Table 2.3: Comparison of sample requirements for EMPD, with a small ($p_0 = 10$) or a medium size ($p_0 = 100$) marker panel, to achieve at least 95% relative accuracy. Accuracies were calculated using SVM leave-one-out cross-validation. $N/2$ denotes the total number of samples per group in the datasets.

2.4 Marker panel variability

We also investigated the overlap of marker panels across 30 runs of random subsampling. The good performance of EMPD suggested that there are many informative genes and that prediction can be based on different combinations of them. Especially for small n_0 the marker panels hardly overlap at all (Fig. 2.8).

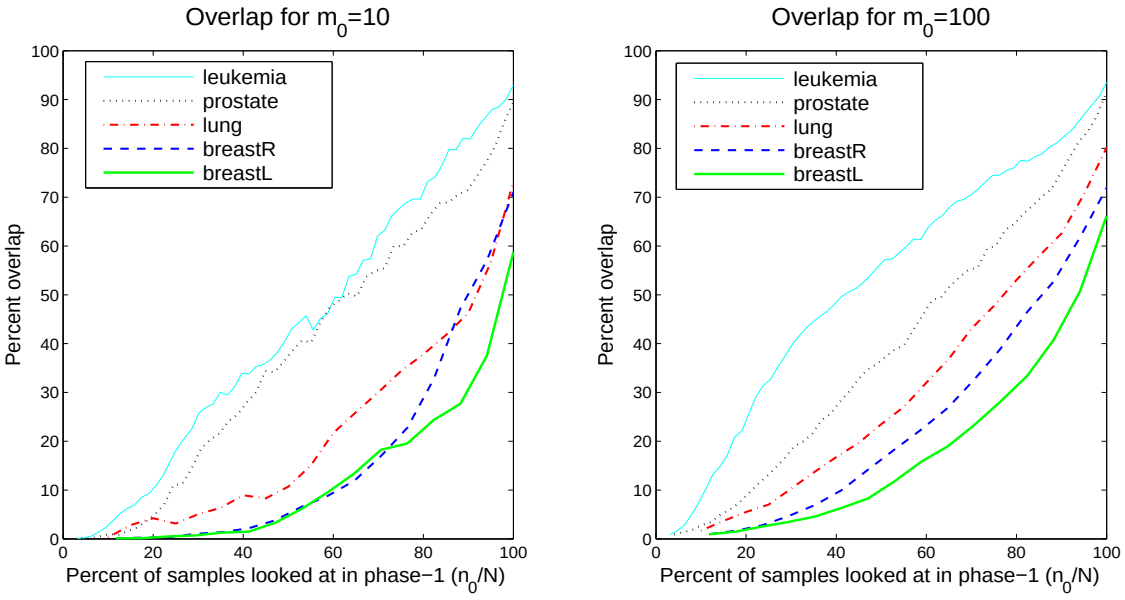

Figure 2.8: Mean pairwise overlap of the marker panels in the 30 subsamplings for each n_0 and $p_0 = \{10, 100\}$

2.5 Discussion

In this chapter, I proposed 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 the phase-1 data. From now on, this marker panel is used to screen a large patient pool (phase-2). Furthermore, I introduced 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 I found that in phase-1 as little as 16 patients per group were sufficient to identify a panel of 10 marker genes. For a marker panel of 100 genes, not more than 10 patients per group were needed. The early decision on the marker panel compromised the final performance of the diagnostic classification only marginally. I showed that there was an inverse relationship between the number of samples in phase-1 and the size of the marker panel. Using more samples in phase-1 facilitated the identification of a more reliable set of markers. Therefore, fewer markers were 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.

I demonstrated that EMPD is a feasible design for cost efficient clinical studies based on gene expression levels. Material, production, and handling costs are 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 I obtained 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. The observation that finding marker genes for classification is easy and does not require many patients, suggests that there are 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) indicated several thousand differentially expressed genes in all four studies, too. 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 the 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 chapter is purely descriptive, and the findings only apply to the five datasets shown. However, since the results in all five studies were consistent, I 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. The evaluation procedure does not allow 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 power-law extrapolation that can be extended to EMPD as well.