Leukemia (2004) 18, 1270–1277

www.nature.com/leu

Prediction of immunophenotype, treatment response, and relapse in childhood acute lymphoblastic leukemia using DNA microarrays

H Willenbrock[1,4], AS Juncker[1,4], K Schmiegelow[2], S Knudsen[1] and LP Ryder[3]

[1]*Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark;* [2]*The Pediatric Clinic II, The University Hospital, Rigshospitalet, Copenhagen, Denmark; and* [3]*Department of Clinical Immunology, Tissue Typing Laboratory, The University Hospital, Rigshospitalet, Copenhagen, Denmark*

Gene expression profiling is a promising tool for classification of pediatric acute lymphoblastic leukemia (ALL). We analyzed the gene expression at the time of diagnosis for 45 Danish children with ALL. The prediction of 5-year event-free survival or relapse after treatment by NOPHO-ALL92 or 2000 protocols resulted in a classification accuracy of 78% and a Matthew's correlation coefficient of 0.59 independently of immunophenotypes. The sensitivity and specificity for prediction of relapse were 87% and 69% respectively. Prediction of high *vs* low levels of the minimal residual disease (MRD) on day 29 ($\geqslant$0.1% or $\leqslant$0.01%) resulted in an accuracy of 100% for precursor-B samples. The classification accuracy of precursor-B- *vs* T-lineage immunophenotypes was 100% even in samples with as little as 10% leukemic blast cells, and the immunophenotype classifier constructed in this study was able to classify 131 of 132 samples from a previous study correctly. Our study indicates that the Affymetrix Focus Array GeneChip may be used without loss of classification performance compared to previous studies using the far more extensive U133A + B GeneChip set. Further studies should focus on prediction of MRD, as this prediction would relate strongly to long-term outcome and could thus determine the intensity of induction therapy.

Leukemia (2004) **18,** 1270–1277. doi:10.1038/sj.leu.2403392
Published online 20 May 2004
Keywords: acute lymphoblastic leukemia; gene expression profiling; DNA microarray; classification; relapse prediction; MRD prediction

Introduction

In the Nordic countries, acute lymphoblastic leukemia (ALL) has an annual incidence of approximately 3.9 per 100 000 children.[1] The diagnosis of ALL is currently based on morphological, immunophenotypic, and cytogenetic analyses of a bone marrow sample, as well as clinical examinations. Based on biological and clinical features, the patients are assigned to risk group adapted therapy, and the 5-year event-free survival (EFS) rate has increased to more than 75% within the last decade.[2,3]

Gene expression profiling of childhood ALL cases has previously shown great promise in diagnosing and risk classification of ALL. Several studies have shown that it is possible to distinguish between the two major immunophenotypes, precursor-B- (preB-) and T-lineage ALL, when applying classification methods based on the genetic profile.[4–6] Furthermore, it has been reported possible to predict relapse as well as development of secondary acute myeloid leukemia within certain subgroups of ALL with 97–100% accuracy.[6]

Correspondence: Dr S Knudsen, Center for Biological Sequence Analysis, Technical University of Denmark, 2800 Lyngby, Denmark; Fax: +45 45 93 15 85; E-mail: steen@cbs.dtu.dk
[4]These two authors contributed equally to this work
Received 12 January 2004; accepted 25 March 2004; Published online 20 May 2004

The aim of this study was to further explore the potential for prediction of relapse and classification of ALL subtypes. We present an attempt for prediction of long-term outcome and treatment response using microarray analysis of diagnostic bone marrow samples from children with ALL who have been treated according to the NOPHO-ALL92 protocol.[2] The long-term outcome was here assessed by either continuous complete remission (CCR), as judged by 5-year EFS, or relapse in the same period. Treatment response was predicted according to either low or high ($\leqslant$0.01% or $\geqslant$0.1%) minimal residual disease (MRD) measured on day 29 of treatment.[7] Furthermore, we attempted the classification of the two major prognostically relevant immunophenotypes, preB- and T-lineage ALL, also for samples with less than 75% leukemic blasts. In this study, we based our gene expression analysis on the Affymetrix Focus Array GeneChip consisting of 8763 well-characterized human genes from the Affymetrix U133A GeneChip. Further material and raw data may be found at http://www.cbs.dtu.dk/~hanni/ALL.

Methods

Patients and material

The study material included children with ALL for whom cryopreserved mononuclear bone marrow cells had been stored at the time of diagnosis, and who were diagnosed between January 1, 1992 and April 1, 2003 and treated at the University Hospital, Rigshospitalet, in Denmark according to the NOPHO ALL-92 protocol or the NOPHO ALL-2000 protocol. Patients who failed to achieve remission, died during induction therapy or in remission, or who developed a second neoplasm were excluded from the study. Criteria for classification as standard (SR), intermediate (IR), high (HR), and very high risk (VHR) have been published previously.[2]

The material included 45 ALL patients, 15 girls and 30 boys with a median age of 8.3 years (range 1–15) at the time of diagnosis. Written consent was obtained for all included patients. The 38 patients diagnosed between January 1992 and June 2001 and treated according to the NOPHO ALL-92 included six cases of SR-ALL, 10 cases of IR-ALL, and 22 cases of HR/VHR-ALL. For these patients, induction, consolidation, and maintenance therapy has been detailed elsewhere.[2] Of the 38 patients, 13 developed a relapse, while 21 patients had a 5-year EFS on April 1, 2003 and are referred to as patients with continuous complete remission (CCR) (Table 1). The remaining four patients as well as the seven patients treated according to the NOPHO ALL-2000 protocol have been followed less than 5 years from the date of diagnosis (Table 1). Among the patients for whom the day-29 MRD level measurements were available, 24 patients had MRD levels $\geqslant$0.1% (high MRD) and 11 patients

Table 1 ALL patients selected for our study according to 5-year outcome and immunophenotype

	Immunophenotype		
Outcome	*preB-lineage*	*T-lineage*	*Total*
Relapse within 5 years	8	5	13
5-year event-free survival	13	8	21
Unknown 5-year outcome	5	6	11
Total	26	19	45

had MRD levels $\leqslant 0.01\%$ (low MRD). During the first 29 days, all these 35 patients received identical therapy except an extra pulse of doxorubicin on day 8 for HR and VHR patients treated according to the NOPHO ALL-92 protocol. For more details about the patients, see http://www.cbs.dtu.dk/~hanni/ALL.

Percentage of leukemic cells in samples

The percentage of leukemic cells present in each patient sample was estimated from data from immunophenotyping of the samples, based on expression of lineage-specific surface antigens like CD19, CD20, and CD3. Out of all 45 samples, 11 had a leukemic blast percentage of $<75\%$ (eight preB- and three T-lineage patients). Among these samples, one had 10%, four had 30–55% while the remaining six samples had 60–70% leukemic blasts.

RNA amplification and application to microarrays

Total RNA was purified from cryopreserved mononuclear cells using the ToTALLY RNA™ Kit (Ambion). For mRNA amplification, the MessageAmp™ aRNA Kit (Ambion) was applied, with the exception of the cDNA purification, which was made according to the Affymetrix clean-up protocol. All remaining steps were made according to the Affymetrix protocol. The final concentration was adjusted for the starting amount of total RNA, and 10 μg of aRNA (or less if 10 μg of aRNA had not been obtained from the amplification step) was fragmented. Hybridization cocktails for Midi array format were prepared and samples were hybridized to Affymetrix Focus Array GeneChips for 15–17 h and subsequently washed and stained with R-phycoerythrin–streptavidin using the Midi_euk2v3 fluidics protocol. Finally, the GeneChips were scanned using the Agilent GeneArray® Scanner to determine the fluorescence intensity for each probe on the chip. Intensities for all probes were saved in a 'CEL file' for subsequent analysis.

Initial data treatment and statistical analysis

The R statistical software[8] was used for the initial data treatment, statistical analysis, and for classification.

Raw probe intensities were normalized using qspline, a nonlinear normalization method.[9] Gene expression indices were calculated using the method of Li and Wong[10,11] with outlier detection using only perfect matches, and background correction using a method implemented in the Robust Multichip Analysis method for calculating expression indices.[12] Unsupervised analysis was performed by hierarchical cluster analysis of patients using Euclidian or vector angle distances.

Feature (gene) selection for the classification was carried out by ranking genes according to their *P*-values in Welsh *t*-test. Further dimension reduction was performed by principal component analysis on a number of selected genes.[13] The classification was carried out on a training set consisting of 2/3 of the data samples randomly selected. Various classification methods were applied and evaluated: *k*-nearest neighbor (KNN),[13,14] nearest centroid (NC),[15] maximum likelihood (ML),[16] nearest shrunken centroid (NSC),[17] linear discriminant analysis (LDA),[15,18] and support vector machines (SVM).[19] Classifier performance was evaluated by leave-one-out cross-validation (LOOCV) and classification accuracy as well as Matthews[20] correlation coefficient (CC). For further information on the applied statistical data treatment, see http://www.cbs.dtu.dk/~hanni/ALL.

Prediction of immunophenotype

Classification of the preB- and T-lineage immunophenotypes was based on the 34 patients with $\geqslant 75\%$ leukemic cells in the samples (18 had preB- and 16 T-lineage). For training, 2/3 of the data set (23 samples) was randomly chosen to comprise a training set. Only three simple classification methods were applied: KNN, NC, and ML algorithms. For the choice of the number of 'general class discriminatory genes', the 50 top ranking genes were evaluated with respect to their appearance in each of the 23 LOOCV *t*-test to determine the genes present in all top 50 LOOCV *t*-tests. These genes were used for training and testing of one optimal classifier for each method. These optimal classifiers were also used for testing of the data set consisting of the 11 patient samples with less than 75% leukemic blast cells.

Prediction of immunophenotype for samples from a previous published study

The raw microarray data (CEL files) for the 132 samples from the study of Ross *et al*[4] were obtained as test samples for our immunophenotype classifier. First, the CEL files were normalized against each other and expression indices were calculated by the same procedure as used for our own chip data. Each U133 GeneChip sample was reduced to probe sets included on the Focus GeneChips and each sample was subsequently normalized – one at a time – against all of the Focus GeneChips from our study, using the qspline normalization method.[9] These data were applied as a test set for our immunophenotype classifier.

Prediction of relapse

Patients with either relapse or CCR as well as $\geqslant 75\%$ of leukemic cells in the sample were used for classification of relapse. Here, 10 relapsed patients (six preB- and four T-lineage) and 18 CCR patients (10 preB- and eight T-lineage) were included. The number of input genes was varied from 2 to 150, and the number of principal components on selected genes was varied from 2 to 12. Several classification methods were applied: KNN, ML, NC, NSC, LDA, and SVM. Random sampling, training on 2/3 of the data set (19 samples), and testing on the independent samples (nine samples) were performed a total of 10 times to ensure that the obtained classification performance was not due to sampling effects. The

30 top ranking genes were evaluated by their presence in at least nine of the 19 LOOCV *t*-tests in at least four of the 10 random samplings to retrieve the 'general class discriminatory genes'.

Prediction of day-29 MRD levels

Patients with available day-29 MRD level measurements as well as $\geqslant$75% of leukemic cells in the sample were used for classification of MRD levels. Here, 18 patients with high day-29 MRD levels (nine preB- and nine T-lineage) and eight patients with low day-29 MRD levels (six preB- and two T-lineage) were included. The same classification and optimization approach was used as for the prediction of relapse.

Results and discussion

Unsupervised analysis of all samples

A hierarchical cluster analysis of all 45 ALL samples based on all 8763 gene expressions (Figure 1) showed that patients with the preB or T immunophenotype generally grouped separately, although a complete separation of the two subtypes was not observed. In three of the five main clusters, preB and T samples were grouped together, except for one T sample (P25). Based on this unsupervised analysis, it seemed as if the most apparent differences in gene expressions between the ALL patients in our study were those determined by the immunophenotype. The same results have previously been obtained by application of unsupervised analysis to ALL microarray data,[5] where preB and T immunophenotypes were identified as the two major subclasses of ALL with almost complete separation.

Regarding the prognostic factors, WBC and age at the time of diagnosis as well as relapse or CCR, no subclustering within the preB or T clusters was observed for the 45 patients.

Classification of immunophenotype

It has previously been reported that the gene expression profiles of preB and T-ALL were easily separable by means of classification,[5,6] and classification of immunophenotypes was also attempted in our study. For all the applied classification methods (KNN, NC, and ML), an accuracy of 100% was obtained both for training (23 samples) and testing the independent test set (11 samples) on the optimal classifier based on the 29 'general class discriminatory genes'. Among these 29 genes (see http://www.cbs.dtu.dk/~hanni/ALL), several were encoding well-known immunophenotype-specific proteins, that is, CD19 and CD3. Moreover, as little as one single gene, CD74, appeared to be enough to distinguish between the two immunophenotypes of ALL.

Prediction of immunophenotype for samples with $<$75% leukemic cells

In two previous extensive studies, all classifications were based on samples with $\geqslant$75% leukemic cells and it had been questioned if their subtype classifier might perform as well on samples with lower levels of leukemic blasts.[4,6] For classification of the 11 samples with less than 75% leukemic blast cells as either preB or T-ALL by our simplified classifier using the 29 'general class discriminatory genes', all samples were classified correctly for the KNN ($k=1$ and 3) and NC classification methods. Thus, our study indicates that the subtype-specific gene expression profile measured in ALL samples was characteristic enough even in samples with less than 75% leukemic cells, and that samples with as little as 10% leukemic blast cells may be classified correctly with respect to immunophenotype. Gene expression analysis might therefore be an improvement of the immunophenotype identification for patients with a small fraction of leukemic cells, where immunophenotyping might be difficult using the current flow cytometric methods.

Prediction of immunophenotype for samples from a previous published study

The results from the classification of preB and T immunophenotypes obtained in our study were evaluated by testing the 132 samples applied onto Affymetrix U133 GeneChips from the study of Ross *et al*.[4] Prediction of immunophenotype using the optimal classifiers designed in our study resulted in very good performance, where the NC method appeared to be superior (Table 2). Only one sample with MLL rearrangements was misclassified of all 132 samples. This subtype of preB ALL was, however, not represented in our data set.

Generally, our results confirmed that the selected 29 'general class discriminatory genes' were not only applicable for prediction of our particular samples but were general for prediction of immunophenotype. Moreover, nine out of the 29 'general class discriminatory genes' were identical to the 100 genes found to be characteristic for distinguishing between preB and T reported by Ross *et al*.[4]

Prediction of relapse

Hierarchical clustering of preB- and T-lineage patients separately as well as for the pooled preB- and T-lineage patients based on all gene expressions did not reveal any clustering into groups with the same outcome when only the 28 relapse or CCR patients with $\geqslant$75% of leukemic cells in the sample were included in the analysis (see http://www.cbs.dtu.dk/~hanni/ALL).

When various classification methods were applied to predict either relapse or CCR of ALL patients, mean CCs in the range of 0.33–0.59 and corresponding accuracies of 0.69–0.78 were obtained for the 19 random samples used for LOOCV training (Table 3). While most methods had an optimal performance using 3–45 gene expressions, LDA and SVM seemed to perform best when using dimension reduced data in the form of two principal components. The nearest centroid classifier had the best performance, $CC = 0.59 \pm 0.18$, with an optimal number of 30 genes (as can be seen in Figure 2), and this method is also the only one that showed a significant classification performance (with an estimated *P*-value of 0.021). Thus, for all other classifiers than NC, there is a statistically significant chance that classification performances matching the obtained CCs might have been obtained with random prior class assignments with a significance level of 5%. Therefore, the NC method seems to be most suitable for the outcome classification problem, and it is also reasonable that a simple method like NC is optimal when only a limited number of samples is available.

Testing of the nine samples in each of the 10 random independent test sets on the LOOCV nearest centroid classifiers

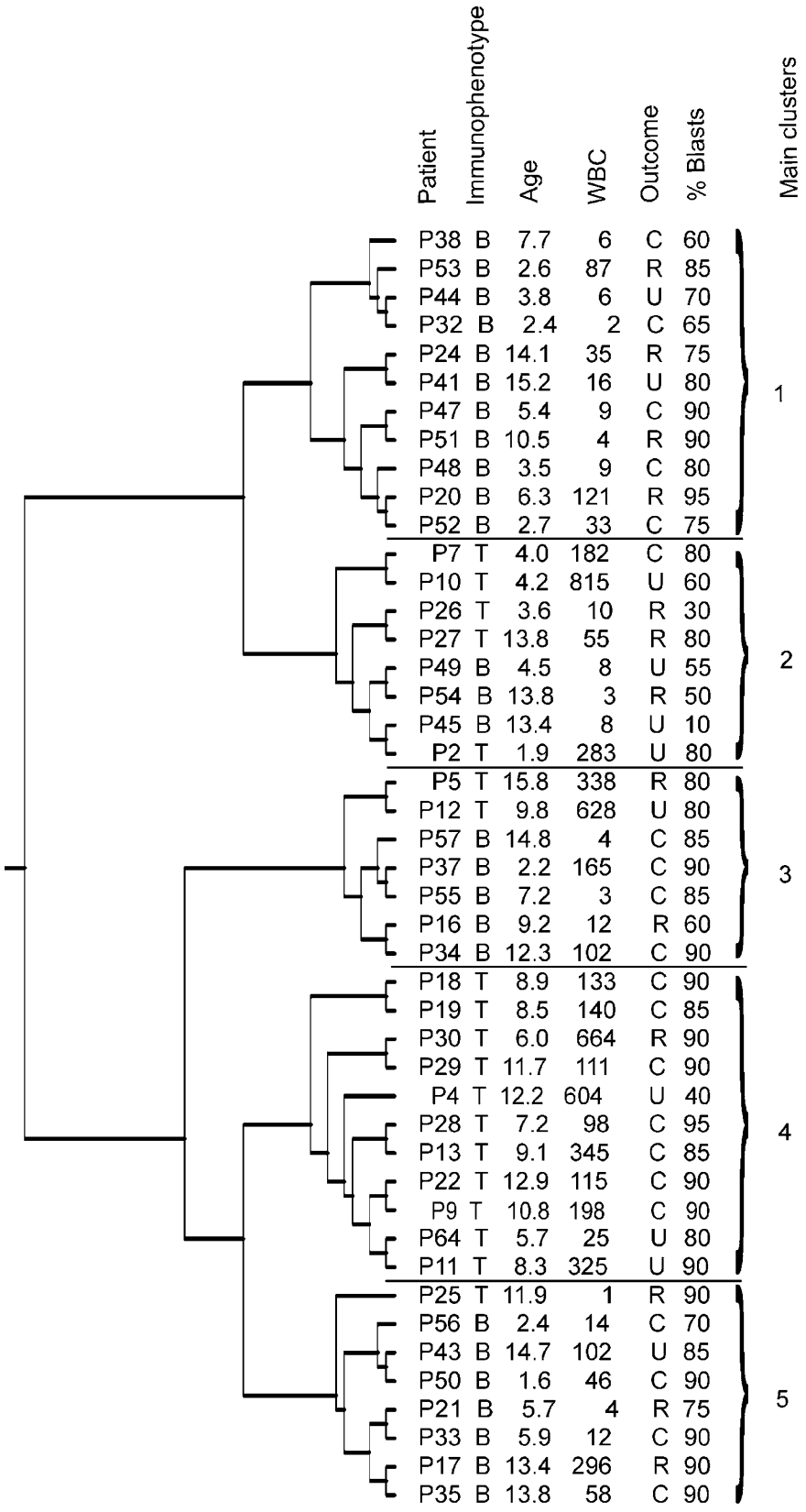

Figure 1 Hierarchical clustering of 45 ALL samples included in the study. From left: patient number, immunophenotype (B: preB-lineage; T: T-lineage), age, WBC (in 10^9/l), 5-year outcome (R: relapse; C: CCR; U: unknown), percentage of leukemic blasts in the sample. The five main clusters are marked.

Table 2 Results from testing of Ross *et al*'s data on the optimal classifier based on the 29 'general class discriminatory genes', trained on 2/3 of the data set with $\geqslant 75\%$ leukemic cells

		preB ALL					
	T-ALL (%)	*E2A-PBX1 (%)*	*Hyperdiploidy (%)*	*MLL (%)*	*BCR-ABL (%)*	*TEL-AML1 (%)*	*Other (%)*
k-nearest neighbor, $k=1$	100	100	100	90	100	100	96.40
k-nearest neighbor, $k=3$	100	100	100	85	100	100	89.30
Nearest centroid	100	100	100	95	100	100	100

Table 3 Prediction of relapse independently of immunophenotype (pooled preB and T-ALL samples)

	Correlation coefficient[a]	*Accuracy*[a]	*Optimal parameters*	*P-value*[a,b]
k-nearest neighbor	0.51	0.77	45 genes, $k^c=3$	0.064
Maximum likelihood	0.44	0.72	4 genes	0.074
Nearest centroid	0.59	0.78	30 genes	0.021
Nearest shrunken centroid	0.47	0.72	3 genes	0.054
Linear discriminant analysis	0.41	0.71	2 PC (based on 4 genes)	0.082
Support vector machine	0.33	0.69	2 PC (based on 30 genes), $c^d=3$	0.074

[a]The values are the average of 10 random samples from LOOCV training of classifiers trained on 2/3 of the data set.
[b]As determined by a permutation test.
[c]Specific parameter for *k*-nearest neighbor.
[d]Specific parameter for support vector machines.

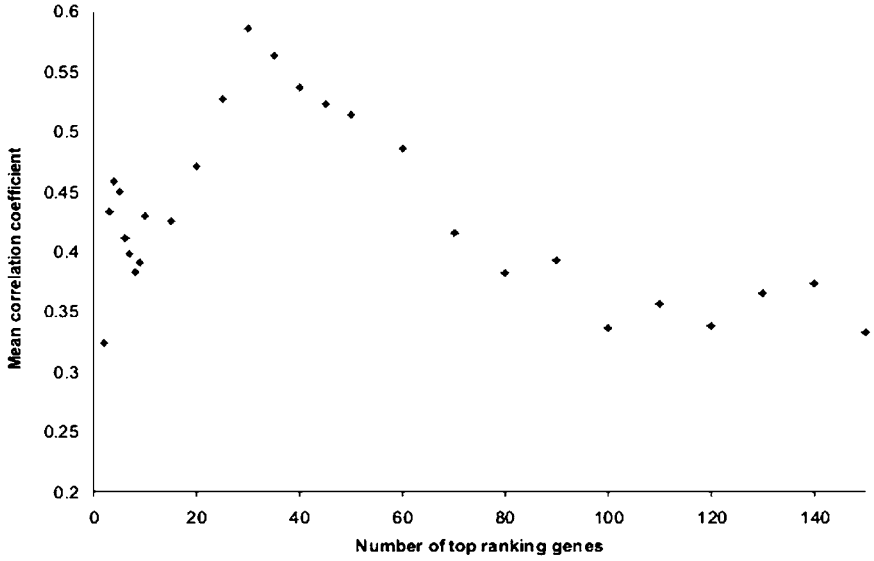

Figure 2 Prediction of relapse. The correlation coefficient (average value for the 10 random samplings) as a function of number of input genes for prediction of outcome (relapse or CCR) for the nearest centroid method.

resulted in a CC of 0.56 ± 0.2 and a corresponding accuracy of $74\%\pm0.11$. This CC and accuracy are thought to be a truer estimate of the classification performance than obtained for LOOCV during training and it is noteworthy that they do not differ significantly.

Interestingly, while the overall accuracy for prediction of outcome was found to be only 74%, the prediction accuracy for the relapse was 87%. The fact that almost all patients with relapse were found among the patients predicted as relapsed based on the diagnostic samples might be important for clinical application of the prediction method, since these patients could have been given an alternative or more intensive treatment. On the other hand, among the patients predicted to be CCR patients, a high percentage is in fact CCR patients (specificity of 92%). Future studies are needed to explore whether this subset can be cured with less intensive therapy. The specificity for patients predicted as relapse patients was only 69%. However, this is a far better overall specificity of relapse prediction than that obtained presently by conventional risk classification criteria such as age, white cell counts, immunophenotype, and cytogenetics.

By evaluation of the 30 top ranking genes, 19 'general class discriminatory genes' were retrieved (see http://www.cbs.dtu.dk/~hanni/ALL). A hierarchical cluster analysis of the 28 patients based on the gene expression of these 19 genes (Figure 3) illustrates that the relapsed patients group together

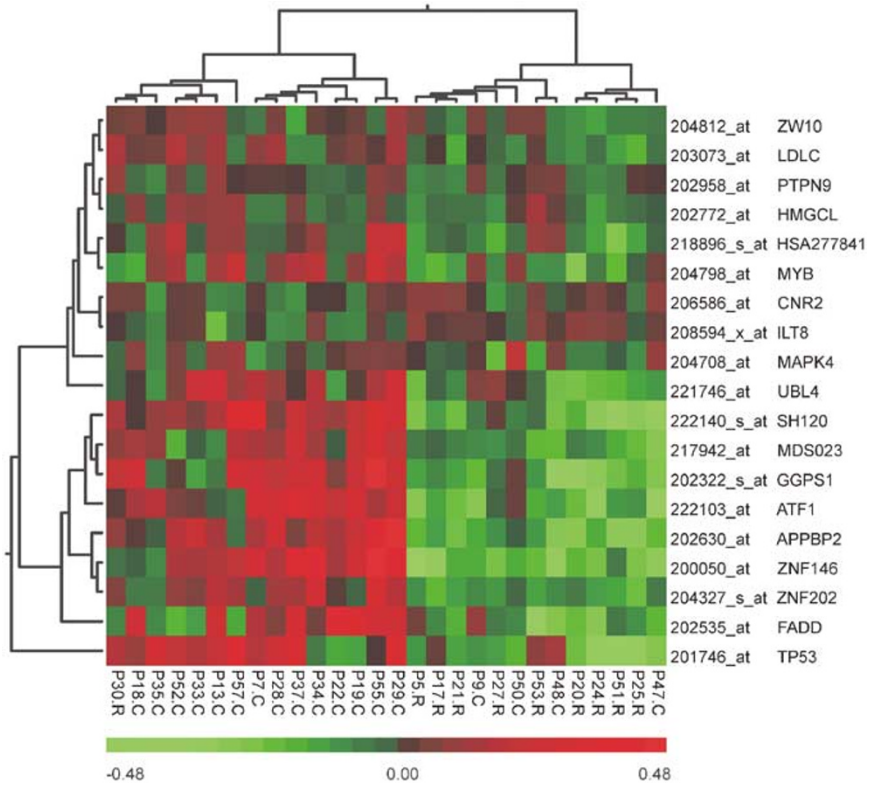

Figure 3 Horizontal: Hierarchical clustering of the 28 ALL patients with known 5-year outcome based on the gene expression levels of the 19 genes. For each patient, the number and outcome, relapse (R) or CCR (C), are given. Vertical: Hierarchical clustering of the 19 genes found to be predictive for long-term outcome based on gene expression levels. The Affymetrix id and gene symbol is given for each gene. The color scale shows the logarithm of the gene expression value relative to the mean logarithmic gene expression for each gene.

with four CCR patients, while only one single relapse patient cluster with the CCR patients. The cluster analysis pattern thus supports the fact that several of the CCR patients are predicted as relapsed patients, while only few relapse patients are predicted as CCR patients.

It has previously been reported possible to predict relapse in certain subgroups of ALL by use of gene expression data[6] with a prediction accuracy of 97% for T-lineage ALL. However, it has later been discovered that this prediction accuracy was overestimated since it was based on LOOCV only during classifier training while the feature selection step had not been included in the LOOCV procedure. The performance was subsequently re-estimated (James R Downing, December 2003) and resulted in a much lower classification accuracy of 73.5% using the top 50 ranked genes in a *t*-test and data pretreatment by Affymetrix MAS 5.0. However, the specificity for relapse cases was only 25% giving a CC of 0.16.

The CCs for prediction of relapse independently of immunophenotype found in the present study (0.59 and 0.56 for the LOOCV training and the independent test sets, respectively) were significantly higher than the CC obtained for T-lineage samples for the re-evaluated data set from Yeoh *et al*,[6] while the prediction accuracy obtained in the present study (78%) was only slightly higher than the re-estimated accuracy obtained by Yeoh *et al*.[6] The better results obtained in our study might partly be due to the different treatments that patients had received in these two studies as well as differences in the period for EFS applied to define patients with CCR, where we defined the minimum period of EFS to be 5 years, while patients with shorter EFS period were included as CCR patients in the study of Yeoh *et al*.[6]

Moreover, it was reported by Yeoh *et al*[6] that it was not possible to predict relapse across subtypes of ALL. However, the results from our study indicated that an at least as good classification performance could be obtained when predicting relapse independently of ALL immunophenotype compared to prediction of relapse for preB and T patients separately (data not shown). However, this may partly be attributed to the fact that a limited number of patients were available for each of these subtype-specific classifiers. Especially, when taking into consideration that there are many subtypes of ALL, we cannot expect to find a common expression profile for relapse for all subtypes. Thus, the low prediction accuracy of clinical outcome is not surprising. The chances for cure for individual patients will reflect the leukemic clone, the host, and the treatment. A number of different leukemia-related biological features such as chromosomal translocations, multiple drug resistance gene activity, and deregulated apoptotic pathways may influence clinical outcome, and their impact may differ between different subsets of ALL. In addition, the strongest prognostic factor is treatment itself. Thus, patients are assigned to different risk groups that are offered different treatment protocols, the bioavailability and disposition of the anticancer agents may differ among patients, and both physician and patient compliance to the treatment protocols may significantly influence the chances for cure. Further improvement of the outcome prediction using DNA microarrays may necessitate analysis of both tumor samples and patient germline samples that allow identification of genetic polymorphisms that influence drug disposition. Such data should be analyzed within biological well-defined subsets of leukemias treated by similar therapeutic strategies.

Prediction of MRD level

Finally, prediction of high and low MRD level on day 29 after treatment initiation was attempted. Since the number of patients with available MRD data was limited, this prediction was only possible for pooled preB and T-ALL patient samples (26 patients) or for preB samples only (15 patients), where only samples with $\geqslant$75% leukemic blasts were included in both cases.

For the pooled preB and T patients, a very low CC was obtained at LOOCV training on 2/3 of the data set (-0.05 to 0.23). On the contrary, for the preB samples only, a classification accuracy of 100% could be obtained during LOOCV training for the LDA and SVM methods on six PC based on the 120 top ranking genes. However, these promising results could not be tested on an independent test set due to the limited number of samples, and when testing on the samples with $<$75% leukemic blasts four out of the six preB samples with available MRD data were predicted correctly (66.7%).

If the MRD classifier was in fact as good as the results from the LOOCV training indicate, it would be highly useful in clinical settings for choice of induction therapy. Thereby, it would be possible to predict the treatment response on day 29 already at the time of treatment initiation, and for the patients with predicted high MRD, an alternative or more intensive therapy could subsequently be given. Another advantage of the MRD prediction is that all patients have received almost identical treatment during the first 29 days, which makes the classification results more easily interpretable and more general compared to the prediction of relapse where treatment during the first 5 years from diagnosis varied among the patients.

Microarray platform

All the analyses performed and the results obtained indicate that the use of the limited Focus Array platform does not result in a loss in classification performance compared to previous studies using the far more extensive U133A + B GeneChip set from Affymetrix for immunophenotype classification. On the contrary, by using the Focus Array, only the most validated genes are included in the analysis and much of the potential noise from probes against possible nonexisting human genes is therefore avoided. While this chip is less complex and easier to interpret than other chips on the market, it has other advantages too. It is cheaper, requires smaller amounts of sample, and is faster to run. Thus, the use of these chips for clinical gene expression profiling seems promising.

Conclusion

Altogether, our results indicate that gene expression analysis using DNA microarrays is a promising tool for prediction of relapse or treatment response in childhood ALL patients. Moreover, our immunophenotype classifier was able to classify correctly all but one sample from a previous independent study and thus, this technology shows potential for future clinical multicenter studies.

Acknowledgements

We thank staff members from the Section of Clinical Hematology and Oncology, Rigshospitalet, laboratory technicians from the Department of Clinical Immunology, Rigshospitalet and people at Center for Biological Sequence Analysis, Technical University of Denmark for their assistance. This work was supported financially by Knud Veilskov's Foundation, Ellen and Aage Fausbølls Health Foundation of 1975, Holger and Inez Petersens Foundation, Gangsted Foundation, Vilhelm Pedersens Foundation, Danish National Research Foundation, Danish Biotechnology Instrument Centre, Danish Center for Scientific Computing, Novo Nordisk, Novozymes, Carlsberg Foundation, The Danish Cancer Society (Grant Nos. 99 144 10 9132, 94-100-28, and 96-100-07), The Danish Cancer League, The Edith & Søren Kiilerich Hansen Family Foundation, The Emil and Inger Hertz Foundation, The Kornerup Foundation, The Lundbeck Foundation, The Medical Research Council in Denmark (Grant No. 9401011), and The Queen Louise's Children's Hospital Foundation.

References

1 Hjalgrim LL, Rostgaard K, Schmiegelow K, Soderhall S, Kolmannskog S, Vettenranta K *et al*. Age- and sex-specific incidence of childhood leukemia by immunophenotype in the Nordic countries. *J Natl Cancer Inst* 2003; **95**: 1539–1544.
2 Gustafsson G, Schmiegelow K, Forestier E, Clausen N, Glomstein A, Jonmundsson G *et al*. Improving outcome through two decades in childhood ALL in the Nordic countries: the impact of high-dose methotrexate in the reduction of CNS irradiation. Nordic Society of Pediatric Haematology and Oncology (NOPHO). *Leukemia* 2000; **14**: 2267–2275.
3 Pui CH, Relling MV, Campana D, Evans WE. Childhood acute lymphoblastic leukemia. *Rev Clin Exp Hematol* 2002; **6**: 161–180, 200–202.
4 Ross ME, Zhou X, Song G, Shurtleff SA, Girtman K, Williams WK *et al*. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood* 2003; **102**: 2951–2959.
5 Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP *et al*. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; **286**: 531–537.
6 Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R *et al*. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 2002; **1**: 133–134.
7 Nyvold C, Madsen HO, Ryder LP, Seyfarth J, Svejgaard A, Clausen N, *et al*., Nordic Society for Pediatric Hematology and Oncology. Precise quantification of minimal residual disease at day 29 allows identification of children with acute lymphoblastic leukemia and an excellent outcome. *Blood* 2002; **99**: 1253–1258.
8 Ihaka R, Gentleman R. R: a language for data analysis and graphics. *J Comp Graph Stat* 1996; **5**: 299–314.
9 Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielsen HB *et al*. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol* 2002; **3**: 1–16.
10 Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA* 2001; **98**: 31–36.
11 Li C, Wong WH. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* 2001; **2**: 1–11.
12 Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003; **4**: 249–264.
13 Knudsen S (ed). *A Biologist's Guide to Analysis of DNA Microarray Data*. New York: Wiley-Interscience, 2002.
14 Dudoit S, Fridlyand J. Introduction to classification in microarrray experiments. In: Berrar DP, Dubitzky W, Granzow M (eds). *A Practical Approach to Microarray Data Analysis*. Dordrecht: Kluwer Academic Publishers, 2003.
15 Dudoit S, Fridlyand J. Classification in microarray experiments. In: Speed T (ed). *Statistical Analysis of Gene Expression Microarray Data*, Interdisciplinary Statistics. Boca Raton, FL: CRC Press, 2003.

16 Dyrskjøt L, Thykjaer T, Kruhoffer M, Jensen JL, Marcussen N, Hamilton-Dutoit S *et al*. Identifying distinct classes of bladder carcinoma using microarrays. *Nat Genet* 2003; **33**: 90–96.
17 Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA* 2002; **99**: 6567–6572.
18 Conradsen K (ed). Diskriminantanalyse. In: *En introduktion til statistik,* Bind 2, 5th edn. Lyngby: IMM, 2002, pp 309–342.
19 Cortes C, Vapnik V. Support-vector network. *Mach Learn* 1995; **20**: 1–25.
20 Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 1975; **405**: 442–451.