Leukemia (2004) 18, 63–71

www.nature.com/leu

Pediatric acute lymphoblastic leukemia (ALL) gene expression signatures classify an independent cohort of adult ALL patients

A Kohlmann[1], C Schoch[1], S Schnittger[1], M Dugas[2], W Hiddemann[1], W Kern[1] and T Haferlach[1]

[1]*Laboratory for Leukemia Diagnostics, Department of Internal Medicine III, Ludwig-Maximilians-University, Munich, Germany; and* [2]*Department of Medical Informatics, Biometrics and Epidemiology, Ludwig-Maximilians-University, Munich, Germany*

Recent reports support a possible future application of gene expression profiling for the diagnosis of leukemias. However, the robustness of subtype-specific gene expression signatures has to be proven on independent patient samples. Here, we present gene expression data of 34 adult acute lymphoblastic leukemia (ALL) patients (Affymetrix U133A microarrays). Support Vector Machines (SVMs) were applied to stratify our samples based on given gene lists reported to predict MLL, BCR-ABL, and T-ALL, as well as MLL and non-MLL gene rearrangement positive pediatric ALL. In addition, seven other B-precursor ALL cases not bearing t(9;22) or t(11q23)/MLL chromosomal aberrations were analyzed. Using top differentially expressed genes, hierarchical cluster and principal component analyses demonstrate that the genetically more heterogeneous B-precursor ALL samples intercalate with BCR-ABL-positive cases, but were clearly distinct from T-ALL and MLL profiles. Similar expression signatures were observed for both heterogeneous B-precursor ALL and for BCR-ABL-positive cases. As an unrelated laboratory, we demonstrate that gene signatures defined for childhood ALL were also capable of stratifying distinct subtypes in our cohort of adult ALL patients. As such, previously reported gene expression patterns identified by microarray technology are validated and confirmed on truly independent leukemia patient samples.
Leukemia (2004) **18,** 63–71. doi:10.1038/sj.leu.2403167
Published online 30 October 2003
Keywords: acute lymphoblastic leukemia; gene expression; microarray; diagnosis

Introduction

Acute lymphoblastic leukemias (ALLs) and acute myeloid leukemias (AMLs) are both heterogeneous diseases.[1] Several subtypes can be discriminated based on cytomorphology, immunophenotype, and recurrent chromosomal aberrations. Inspired by the pivotal discrimination of unselected ALL and AML samples based on their gene expression signatures,[2] recent microarray studies demonstrated the close correlation of immunophenotypical characteristics and chromosomal aberrations in prognostically important leukemia subtypes to distinct gene expression patterns.[3–7] The specific acute leukemia subtypes can be classified by gene expression signatures with exceedingly high accuracies. However, those findings are based on a limited number of patient samples or training and testing sets, respectively. More importantly, discriminative genes were validated based on expression profiles generated in one specific setting of an individual laboratory. In order to become generally accepted as an additional diagnostic method, the robustness of subtype-specific gene expression signatures for leukemia subclassification has to be proven on independent patient samples. We therefore asked the question whether differences in gene expression identified by other groups can also be used to differentiate our patient cohort.

Here, we analyzed the gene expression patterns (Affymetrix U133A microarrays) of our own cohort of 34 adult leukemia patients comprising precursor B-ALLs with MLL gene rearrangements ($n=10$), and translocation t(9;22) ($n=15$) as well as precursor T-ALLs ($n=9$). We applied the diagnostic compositions of candidate genes as reported by Yeoh *et al*[7] and Armstrong *et al*,[3] respectively, to stratify our cases. Thus, the aim of this study was to validate both the reported differentially expressed genes and to evaluate the applicability of pediatric gene expression signatures to predict adult ALL subtypes.

Secondly, we analyzed seven genetically more heterogeneous adult B-precursor ALL cases not bearing the above-mentioned chromosomal aberrations. Following a similar strategy, which was reported by Ferrando *et al*[8] for discovering novel oncogenes in T-ALL, the more heterogeneous B-precursor ALLs were projected into an ALL subtype relevant gene space.

Materials and methods

Patient samples

This study included bone marrow samples from $n=41$ adult ALL patients at diagnosis representing three distinct ALL subtypes MLL, BCR-ABL, and T-ALL, as well as heterogeneous B-precursor ALL cases (supplemental Table 1). All samples were sent between May 1999 and July 2002 for reference diagnostics to our laboratory and registered in our leukemia database.[9] Samples were received either locally or by overnight mail. Prior to therapy, all patients gave their informed consent for participation in the current evaluation after having been advised about the purpose and investigational nature of the study as well as of potential risks. The study design adhered to the declaration of Helsinki and was approved by the ethics committees of the participating institutions prior to its initiation. The diagnosis was performed by a combination of cytomorphology, cytogenetics, fluorescence *in situ* hybridization (FISH), multiparameter-immunophenotyping, and molecular genetics. A more detailed description of patient characteristics and the routine diagnostic procedures is included as supporting online information.

Microarray experiments and SVM classification

Microarray analyses were performed as previously described.[5,6] In order to achieve comparability of differing sets of microarray expression data, the different Affymetrix U95A chip design and U133A chip design probeset information was matched. Briefly, we extracted and combined the significant U95Av2 probesets specific for MLL, BCR-ABL, and T-ALL subtypes as depicted in the respective publications. Unique U95Av2 probesets were then functionally annotated using the November 11, 2002

Correspondence: A Kohlmann, Laboratory for Leukemia Diagnostics, Department of Internal Medicine III, Ludwig-Maximilians-University – Grosshadern, Marchioninistr. 15, 81377 Munich, Germany; Fax: 49 89 7095 4971
Received 8 April 2003; accepted 3 September 2003; Published online 30 October 2003

NetAffx™ Analysis Center descriptions.[10] Next, we determined for those unique U95Av2 probesets their corresponding U133A counterparts. Genes represented on U95Av2 microarrays are also represented on U133A microarrays. However, due to the improvement of oligonucleotide selection for the U133A array design and the dynamic nature of public databases, probesets of different array designs are not identical. In order to identify the names of the probesets that are most closely related to another, Affymetrix has made comparison spreadsheets available (www.affymetrix.com). We therefore applied a stringent search strategy using the 'Human Genome U95 to Human Genome U133 Best Match comparison spreadsheet'. This search resulted in best-match U133A counterparts for the U95Av2 probesets, which were chosen for the following statistical analyses.

The classification accuracy of a given gene list for a set of microarray experiments was estimated using Support Vector Machines (SVMs) as supervised learning technique. In general, SVMs are trained on a subset of the data, in our study with prior knowledge using previously reported discriminative gene lists for the respective leukemia subtypes, and then this trained model is employed to assign new samples to known groups from a second and different data set. In our approach, the apparent accuracy, that is, the overall rate of correct predictions of the complete data set, was estimated by 10-fold crossvalidation. This means that the data set was divided into 10 approximately equally sized subsets; a SVM model was trained for nine subsets and predictions were generated for the remaining subset. This training and prediction process was repeated 10 times to include predictions for each subset.

Subsequently the data set was split into a training set, consisting of two-thirds of the samples, and a test set with the remaining one-third. The apparent accuracy for the training set was estimated by 10-fold crossvalidation (analogous to apparent accuracy for complete set). A SVM model of the training set was built to predict diagnosis in the independent test set, thereby estimating the true accuracy of the prediction model. This prediction approach was applied both for overall classification (multiclass) and binary classification (diagnosis X $\Rightarrow$ yes or no). For the latter, sensitivity and specificity were calculated:

$$\begin{aligned}\text{Sensitivity} &= (\text{number of positive samples predicted})/\\ &\quad (\text{number of true positives})\\ \text{Specificity} &= (\text{number of negative samples predicted})/\\ &\quad (\text{number of true negatives})\end{aligned}$$

More detailed information on U95A–U133A microarray probeset match strategy, applied statistical methods for data analysis, and classification, as well as raw gene expression intensities of diagnostic markers for download is included as supporting online information.

Supervised identification of differentially expressed genes

To identify the genes whose expression patterns best distinguished among T-ALL, MLL, and BCR-ABL-positive cases in our data series, we applied the SAM software program.[11] Affymetrix U133A signal intensities were transformed as previously described and subsequently inputted into the software.[5] A stringent cutoff for significance (tuning parameter delta) for <1 false-positive rated gene was chosen. A complete list of identified genes including the scaled microarray expression data is available in the online section.

Data visualization

To assess the similarity of gene expression patterns, we applied hierarchical cluster analysis and principal component analyses.[12,13] Transformed U133A expression data were analyzed using the GeneMaths 2.01 software from Applied Maths, Belgium (cluster algorithm: Ward; selected coefficient: Euclidean distance).

Results and discussion

Genes identified by Yeoh *et al* and Armstrong *et al* were represented on Affymetrix HG-U95 chip design microarrays. Meanwhile, the newly designed HG-U133A microarray is available and was utilized for this study. According to the manufacturer's information, both oligonucleotide selection (for further details, see technical note on U133A array design, www.affymetrix.com) and analysis software (Microarray Suite 5.0) were improved compared to previous GeneChip microarrays.[14,15] Additionally, sample assessment, handling and storage, differing target labeling protocols, microarray scanner photo multiplier tube settings, and different software algorithms for analysis of primary expression signal intensities (Microarray Suite software versions 4.0 *vs* 5.0) represent parameters that account for influences comparing the results from different laboratories. Taking all those pitfalls into account, we chose a simple strategy to compare different data sets of gene expression intensities. Briefly, according to Yeoh *et al* and Armstrong *et al*, respectively, all important U95 chip design candidate genes to discriminate ALL with MLL gene translocation, t(9;22)-positive ALL (BCR-ABL) and T-ALL were matched to our corresponding U133A probesets. The raw expression data were transformed as described (online supplemental material). Then we aimed at predicting our independent cohort of ALL patients using common machine learning algorithms (SVM) and a 10-fold crossvalidation approach.[16–18]

ALL subtype prediction using St Jude Children's Research Hospital childhood ALL predictors

First, we compared our expression data to available expression profiles ($n=78$) of St Jude Children's Research Hospital childhood ALL samples (http://www.stjuderesearch.org/data/ALL1). Yeoh *et al* had used Affymetrix oligonucleotide microarrays to analyze the pattern of genes expressed in leukemic blasts from 360 pediatric ALL patients. Distinct expression profiles identified each of the prognostically important leukemia subtypes, including T-ALL, E2A-PBX1, BCR-ABL, TEL-AML1, MLL gene rearrangement, and hyperdiploid >50 chromosomes. They selected discriminating genes for the various ALL subtypes using a variety of statistical metrics. We extracted and combined all significant U95Av2 probesets specific for MLL, BCR-ABL, and T-ALL subtypes. Next, we determined for those unique U95Av2 probesets their corresponding U133A counterparts.

The data presented here indicate that the genes reported by Yeoh *et al* can also separate our cohort of adult ALL patient samples. Subgroup prediction using SVM learning algorithms demonstrates the discriminative properties of those candidate genes specific for T-ALL, BCR-ABL, and MLL subtypes in ALL (Table 1). A hierarchical cluster analysis of our adult ALL samples using the preselected subset of genes specific for MLL, BCR-ABL, or T-ALL confirms the capability of separating three ALL subtypes based on distinct expression signatures. As

Table 1 SVM subtype prediction accuracies using Yeoh *et al* list of genes

Subgroups	*Complete set*[a] *Apparent accuracy (%)*[d]	*Training set*[b,h] *Apparent accuracy (%)*[d]	*Test set*[c,h] *True accuracy (%)*[e]	*Sensitivity (%)*[f]	*Specificity (%)*[g]
Overall	94	87	100		
BCR-ABL	97	87	100	100	100
MLL	97	96	100	100	100
T-ALL	100	91	100	100	100

[a]The complete set consisted of 34 samples.
[b]The training set consisted of 23 samples.
[c]The test set consisted of 11 samples.
[d]Apparent accuracy was determined by 10-fold crossvalidation.
[e]True accuracy was determined by class prediction on the blinded test set.
[f]Sensitivity = (the number of positive samples predicted)/(the number of true positives).
[g]Specificity = (the number of negative samples predicted)/(the number of true negatives).
[h]The distribution of cases in the training and test sets are: BCR-ABL (10, 5 cases), MLL (7, 3), T-ALL (6, 3).

visualized in Figure 1, samples of each of the three distinct ALL subtypes cluster together (upper dendrogram). Based on the given preselected gene expression data, the clustering algorithm accurately assigns the ALL samples according to their underlying genetic aberration and immunophenotype, respectively.

ALL subtype prediction using Dana–Farber Cancer Institute ALL predictors

Secondly, we compared our expression data to available expression profiles of Dana–Farber Cancer Institute ALL samples (http://research.dfci.harvard.edu/korsmeyer/MLL.htm). Armstrong *et al* compared the gene expression profiles of leukemic cells from individuals diagnosed with precursor B-ALL bearing an MLL gene rearrangement against those from patients diagnosed with conventional B-precursor ALL that lack this translocation ($n=44$ pediatric ALL patients). They had determined whether there were genes correlated with the presence of an MLL translocation. Here, we applied that set of published genes to distinguish between MLL and non-MLL cases in our cohort of patients. By applying this preselected set of marker genes, we can robustly distinguish between MLL and non-MLL leukemias in our cohort of adult patients with high accuracy (Table 2). As visualized in Figure 2, based on the given preselected U133A microarray gene expression data, the clustering algorithm accurately groups our ALL samples into MLL gene rearrangement positive and MLL gene rearrangement negative cases.

ALL subtype prediction using overlapping MLL-specific predictors

Finally, we can identify overlapping genes specific for MLL and non-MLL in both published data sets and apply this stringent marker selection to stratify MLL and non-MLL patient samples in our own cohort. A substantial number of genes characterizing MLL-positive patient samples are overlapping between Yeoh *et al* and Armstrong *et al* gene lists (see online supplemental section). In our microarray expression data set, leukemia classification using SVM learning algorithms demonstrates the accurate discriminative properties of those MLL-specific candidate genes (Table 3). Again, based on the given preselected gene expression data, the clustering algorithm accurately groups our adult ALL samples into MLL gene rearrangement positive and MLL gene rearrangement negative cases (Figure 3). As such, in both pediatric and adult ALL patient cohorts, MLL gene rearrangement positive and MLL gene rearrangement negative cases can be robustly predicted.

Molecular characterization of heterogeneous B-precursor ALL cases

After obtaining these results, that specific signatures in childhood ALLs are also observed in adult ALLs, we were interested in the expression profiles of heterogeneous B-precursor ALL cases not positive for t(9;22) or t(11q23)/MLL, respectively. Therefore, we analyzed in addition seven B-precursor ALL patients (c-ALL and Pre-B-ALL) to gain new insights into the molecular features of these cases. This additional cohort comprised patients who showed a normal karyotype ($n=2$) or a variety of different karyotype abnormalities ($n=5$) (supplemental Table 1). A detailed description of respective immunophenotypes and karyotypes is available in the online section.

First an unsupervised analysis, that is hierarchical clustering and principal component analysis (PCA) of the complete data set, was performed. However, this analysis did not reveal informative structures (data not shown).

We therefore applied a supervised analysis using the SAM (significance analysis of microarrays) software to identify differentially expressed genes correlated to T-ALL, BCR-ABL, and MLL cases.[11] A selection of the top 510 genes accurately separated the latter three ALL subtypes in a respective principal component analysis (Figure 4a). Next, other B-precursor ALL samples (without BCR-ABL or MLL chromosomal aberrations) were added to the data set and all cases were projected into the space of the 510 leukemia subtype relevant genes. As shown in Figure 4b, the other B-precursor ALL samples (yellow spheres) intercalate with BCR-ABL-positive samples (red spheres). Thus, the other B-precursor ALL shares similar characteristics with BCR-ABL-positive ALLs. This is in line with the definition and subclassification of B-precursor ALL according to EGIL, which based on the immunophenotype distinguishes Pro-B-ALL, common ALL, and Pre-B-ALL.[19] Most importantly, both other ALL cases and BCR-ABL cases belong to the common ALL and Pre-B-ALL groups, and are thus anticipated to have common gene expression profiles.

This finding can also be visualized using the hierarchical clustering technique (Figure 5).[12] As shown in Figure 5, due to inherent similarities in their expression profiles three major branches of the top dendrogram can be observed. MLL and T-ALL samples are accurately grouped. The more heterogeneous

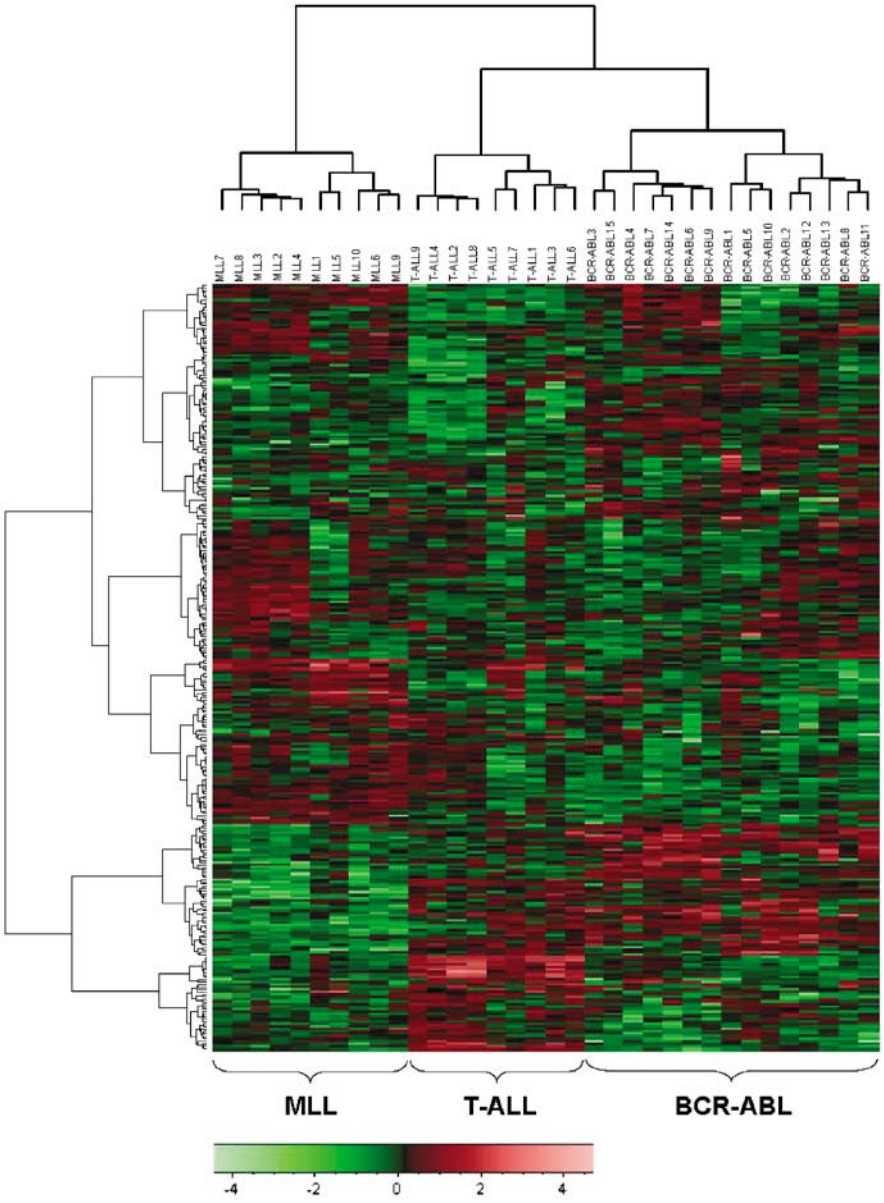

Figure 1 Hierarchical cluster analysis based on U133A microarray expression data of our adult ALL samples (columns) using a subset of genes (rows) identified to classify pediatric ALL samples (Yeoh *et al*). A total of 312 unique best-match U133A probesets were identified to represent the 364 unique U95Av2 probesets according to Yeoh *et al* for the distinction of MLL, BCR-ABL, and T-ALL leukemias. The normalized expression value for each gene is coded by color, with the scale shown at the lower left (s.d. from mean). Red cells indicate high expression and green cells indicate low expression. More detailed information on the genes, that is HGNC approved gene symbol and short functional description, is available as supporting online information.

Table 2 SVM subtype prediction accuracies using Armstrong *et al* list of genes

Subgroups	*Complete set*[a] *Apparent accuracy (%)*[d]	*Training set*[b,h] *Apparent accuracy (%)*[d]	*Test set*[c,h] *True accuracy (%)*[e]	*Sensitivity (%)*[f]	*Specificity (%)*[g]
Overall	100	100	100		
MLL	100	100	100	100	100
non-MLL	100	100	100	100	100

[a]The complete set consisted of 34 samples.
[b]The training set consisted of 23 samples.
[c]The test set consisted of 11 samples.
[d]Apparent accuracy was determined by 10-fold crossvalidation.
[e]True accuracy was determined by class prediction on the blinded test set.
[f]Sensitivity = (the number of positive samples predicted)/(the number of true positives).
[g]Specificity = (the number of negative samples predicted)/(the number of true negatives).
[h]The distribution of cases in the training and test sets are: MLL (7, 3 cases), non-MLL (16, 8).

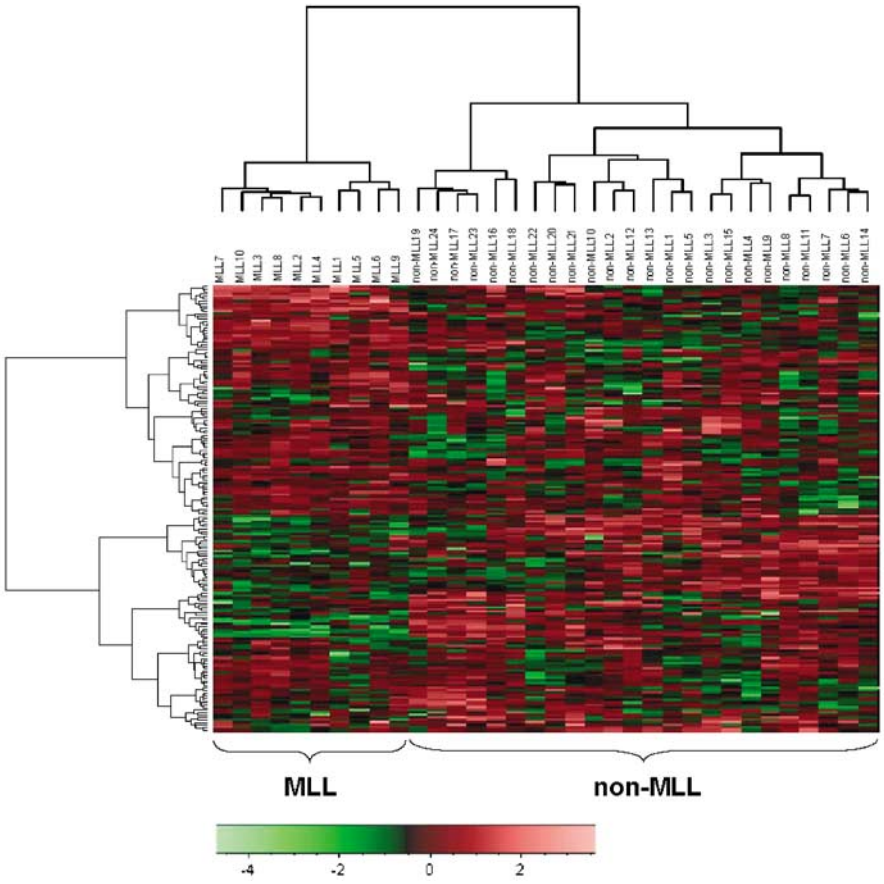

Figure 2 Hierarchical cluster analysis based on U133A microarray expression data of our adult ALL samples (columns) using a subset of genes (rows) identified to classify pediatric ALL with MLL gene rearrangement and non-MLL, respectively (Armstrong *et al*). A total of 182 unique best-match U133A probesets corresponded to 217 identified U95A chip design probesets for the distinction of MLL and non-MLL ALL according to Armstrong *et al*. The normalized expression value for each gene is coded by color, with the scale shown at the lower left (s.d. from mean). Red cells indicate high expression and green cells indicate low expression. More detailed information on the genes, that is HGNC-approved gene symbol and short functional description, is available as supporting online information.

Table 3 SVM subtype prediction accuracies using overlapping MLL-specific genes

Subgroups	*Complete set*[a] *Apparent accuracy (%)*[d]	*Training set*[b,h] *Apparent accuracy (%)*[d]	*Test set*[c,h] *True accuracy (%)*[e]	*Sensitivity (%)*[f]	*Specificity (%)*[g]
Overall	100	100	100		
MLL	100	100	100	100	100
non-MLL	100	100	100	100	100

[a]The complete set consisted of 34 samples.
[b]The training set consisted of 23 samples.
[c]The test set consisted of 11 samples.
[d]Apparent accuracy was determined by 10-fold crossvalidation.
[e]True accuracy was determined by class prediction on the blinded test set.
[f]Sensitivity = (the number of positive samples predicted)/(the number of true positives).
[g]Specificity = (the number of negative samples predicted)/(the number of true negatives).
[h]The distribution of cases in the training and test sets are: MLL (7, 3 cases), non-MLL (16, 8).

B-precursor ALL cases are exclusively distributed in the branch containing all BCR-ABL samples.

Several subtrees in the left dendrogram indicate coexpression of genes for the distinct ALL subtypes. Subtree 1 contains genes overexpressed in T-ALL that have also recently been reported by other microarray studies: *TRB*, *CD3D*, *CD3E*, *CD2*, *CD6*, *MAL*, *LCK*, *ITM2A*, *SH2D1A*.[7,8] A large number of these genes and additional candidates like transmembrane adapters (*LAT*, *TRIM*), further CD3 complex signal transducing members (*CD3G*, *CD3Z*), *CD8A* coreceptor, and *ZAP70* tyrosine kinase could be correlated to a functional role in the class I MHC-restricted T-cell receptor signalosome.[20] As such, the identification of these overexpressed T-ALL-associated candidate genes illustrates the power of gene expression profiling to elucidate complex pathways in a highly parallel manner.

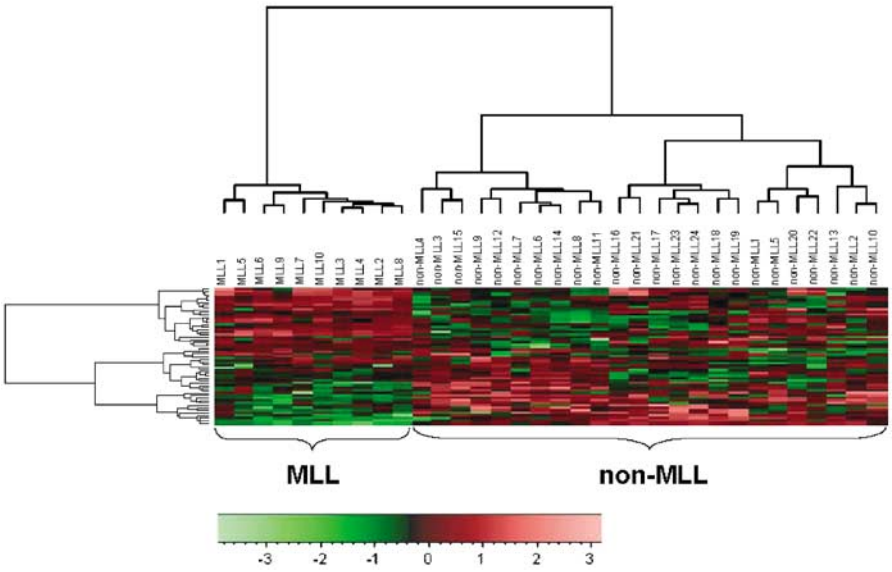

Figure 3 Hierarchical cluster analysis based on U133A microarray expression data of our adult ALL samples (columns) using an overlapping subset of genes (rows) identified to classify pediatric ALL with MLL and non-MLL, respectively. A comparison of both Yeoh *et al* and Armstrong *et al* published gene lists resulted in a number of $n=57$ overlapping U95A chip design probesets reported to be correlated with pediatric ALL-carrying MLL gene aberrations. A total of 55 unique best-match U133A probesets corresponded to those 57 identified U95 chip design probesets. The normalized expression value for each gene is coded by color, with the scale shown at the lower left (s.d. from mean). Red cells indicate high expression and green cells indicate low expression. More detailed information on the genes, that is HGNC-approved gene symbol and short functional description, is available as supporting online information.

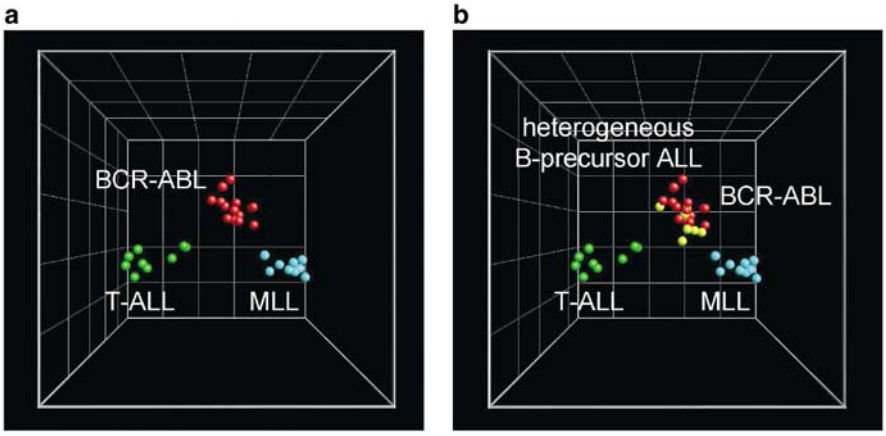

Figure 4 (a) Principal component analysis of T-ALL ($n=9$), MLL ($n=10$), and BCR-ABL ($n=15$) patients. The leukemia samples are plotted in a three-dimensional space using the three components capturing most of the variance in the data set. Each patient sample is represented by a color-coded sphere. Adult ALL cases are accurately separated based on 510 differentially expressed genes identified using the SAM software package. (b) Principal component analysis of T-ALL ($n=9$), MLL ($n=10$), BCR-ABL ($n=15$), and heterogeneous precursor B-lineage B-ALL ($n=7$) patients. The leukemia samples are plotted in a three-dimensional space using the three components capturing most of the variance in the data set. In all, 510 top differentially expressed genes have been identified using the SAM software package in a supervised approach considering the three entities T-ALL, MLL, and BCR-ABL. Each patient sample is represented by a color-coded sphere. Heterogeneous precursor B-lineage ALL samples (yellow spheres) intercalate with BCR-ABL-positive samples (red spheres) when projected in this ALL subtype-specific gene space.

Subtrees 4 and 5 group genes with a high expression in *MLL* gene rearranged positive ALLs. Also, genes that have recently been reported by other microarray studies were *ADAM10*, *BLK*, *CD72*, *CD79A*, *CSPG4*, *HOXA9*, *HOXA10*, *IGHM*, *LGALS1*, *LMO2*, *MBNL*, *MEF2A*, *PPP2R5C*, *PTPRC*, and *VLDLR*.[3,7,21] Candidate genes like *IGHM*, *BLK*, and *CD79A* illustrate the B-lineage characteristics of these cases, and an observed overexpression of *HOXA* cluster members illustrates important components of leukemogenesis driven by *MLL* gene translocations.[3,22]

Subtree 3 mainly contains genes with a functional role in immune response. *BLNK*, *BRDG1*, *CD24*, *MHC2TA*, *CD74*, *HLA-DMA*, *HLA-DMB*, *HLA-DPA1*, *HLA-DRA*, *HLA-DPB1*, *HLA-DQB1*, *HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4*, and *TNFRSF14* demonstrate similar patterns for BCR-ABL, MLL-positive, and the more heterogeneous B-precursor cases. In detail, major components of the class II MHC-restricted antigen presentation machinery are consistently overexpressed compared to T-ALL samples: MHC2TA, interacting with MHC class

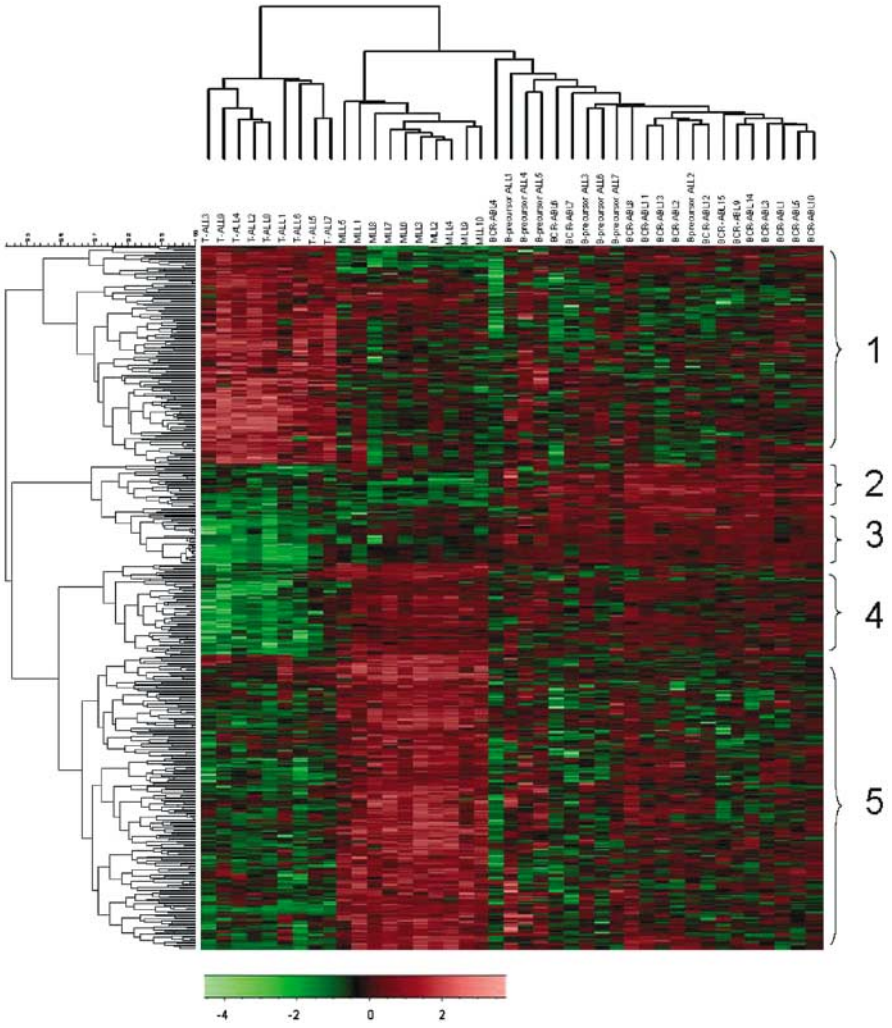

Figure 5 Hierarchical cluster analysis of T-ALL ($n=9$), MLL ($n=10$), BCR-ABL ($n=15$), and heterogeneous precursor B-lineage B-ALL ($n=7$) patients. This analysis is based on 510 differentially expressed genes, which have been identified using the SAM software package in a supervised approach considering the three entities T-ALL, MLL, and BCR-ABL. The normalized expression value for each gene is coded by color, with the scale shown at the lower left (s.d. from mean). Red cells indicate high expression and green cells indicate low expression. A more detailed information on the genes, that is HGNC-approved gene symbol and short functional description, is available as supporting online information.

II as well as HLA-DM and CD74 promotors, is a highly regulated transactivator governing all spatial, temporal, and quantitative aspects of MHC class II expression.[23] The chaperone CD74 (invariant chain) blocks the peptide-binding site of newly synthesized MHC class II molecules by its so-called CLIP fragment.[24] HLA-DM molecules catalyze the exchange of CLIP for antigenic peptides derived from endosomal compartments.

An interesting cluster of genes is organized in subtree 2, which is enlarged in Figure 6. A total of 26 probesets demonstrate similar expression signatures for both BCR-ABL-positive and the more heterogeneous B-precursor cases. All candidate genes are consistently overexpressed in these cases compared to T-ALL and MLL samples. In detail, *LGMN* (legumain), also called asparaginyl endopeptidase (*AEP*), has been reported to be critically involved in the processing of antigens for MHC class II presentation.[25] More recently, a prodrug strategy incorporating a legumain-cleavable peptide substrate onto doxorubicin was developed.[26] A receptor tyrosine kinase activated by collagen, *DDR1* (discoidin domain receptor 1), is represented by three probesets. In a recent report, high-grade primary brain and metastatic brain tumors showed unequivocal, intense *DDR1* expression within the majority of tumor cells.[27] *CD52*, an excellent target for complement-mediated lysis and antibody-dependent cellular cytotoxicity, has been identified by two probesets. Several clinical trials have already been carried out with Alemtuzumab (CAMPATH-1H), a humanized monoclonal antibody directed against the CD52 antigen of lymphocytes.[28] A cytokine-like protein (*C17*), retinoic acid induced gene (*RAI14*), or hypothetical protein *LOC54103* represent further overexpressed genes. However, no functional gene annotation is available yet.

A similar distribution of the adult ALL samples can be observed when these cases were projected into the gene expression space of markers previously reported from Yeoh *et al* to discriminate six distinct pediatric ALL subtypes, that is T-ALL, E2A-PBX1, BCR-ABL, TEL-AML1, MLL, and hyperdiploid leukemias.[7] As anticipated, genetically heterogeneous samples again cluster together with BCR-ABL cases, confirming our previous observation. They do not show up as an independent fourth distinct cluster separated from adult T-ALL, MLL, and

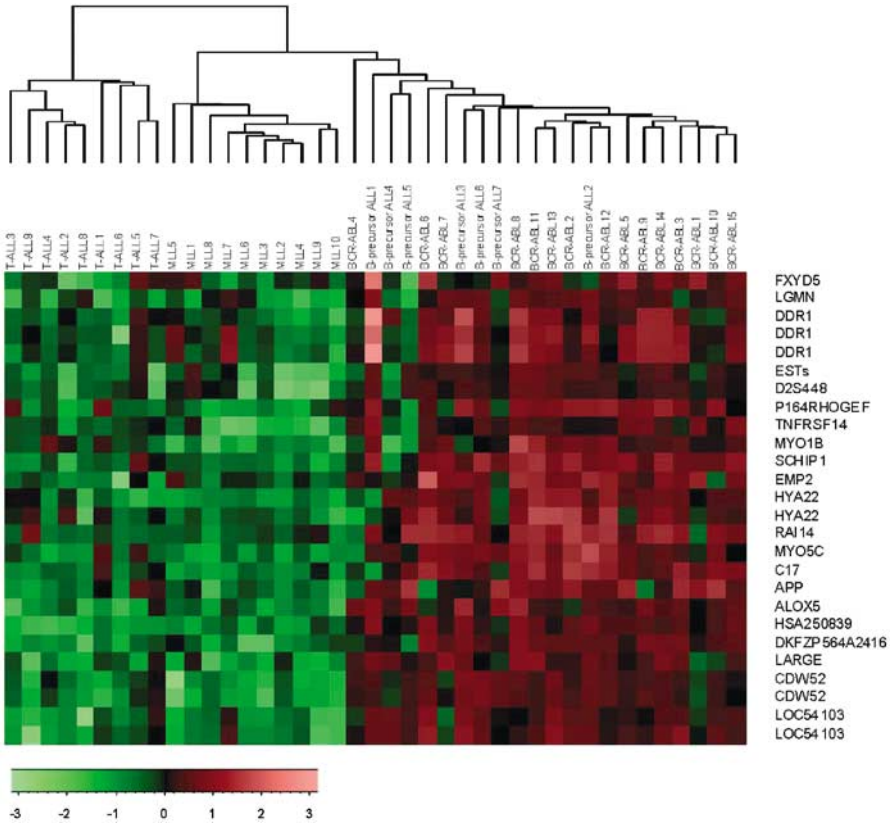

Figure 6 Zoomed image of subtree 2 out of Figure 5. Heterogeneous B-precursor ALL share similar expression patterns with BCR-ABL-positive ALLs. In this subtree, 26 probesets were consistently overexpressed compared to both T-ALL and MLL. The normalized expression value for each gene is coded by color, with the scale shown at the lower left (s.d. from mean). Red cells indicate high expression and green cells indicate low expression.

BCR-ABL-positive leukemias. The observed dendrogram structure of a hierarchical cluster analysis, as well as a 3D plot from a principal component analysis, were added to the supplemental section (Supplemental Figure 1). Thus, signatures, previously reported to correlate with E2A-PBX1, TEL-AML1, and hyperdiploid childhood leukemias could not separate these two groups. This is not unexpected as none of our heterogeneous B-precursor ALL showed any of these genetic characteristics.

Conclusions

Taken together, our observations provide evidence that genes suitable for classification and prediction of childhood ALL are also capable of distinguishing the respective adult ALL subentities. This is a promising finding, as new molecular targets in common genetic subtypes of acute leukemias identified by microarray technology might be common therapeutic targets for both age groups of patients. Previously reported gene expression signatures identified by global genome expression analysis are validated and confirmed on a truly independent patient cohort. Despite influencing parameters such as technical equipment, different sample handling, routine diagnostic procedure, and target preparation for expression analysis by unrelated personnel in an independent diagnostic laboratory, it was demonstrated that ALLs can be classified and predicted based on microarray technology. A prospective validation of diagnostic compositions of candidate genes in clinical trials using less costly, low-density microarrays is warranted. Alternatively, a more minimalized set of discriminative genes may be tested in multiplex-PCR-based assays.

Acknowledgements

This study was supported by a grant from the Deutsche José Carreras Leukämie-Stiftung e.V. (DJCLS-R00/13).

Supplementary Information

Supplementary information is available on the Leukemia website (http://www.nature.com/leu/)

References

1 Jaffe ES, Harris NL, Stein H, Vardiman JW. *World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues*. Lyon: IARC Press, 2001.
2 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.
3 Armstrong SA, Staunton JE, Silverman LB, Pieters R, Den Boer ML, Minden MD *et al*. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* 2002; **30**: 41–47.

4 Kern W, Kohlmann A, Wuchter C, Schnittger S, Schoch C, Mergenthaler S *et al*. Correlation of protein expression and gene expression in acute leukemia. *Cytometry* 2003; **55B**: 29–36.
5 Kohlmann A, Schoch C, Schnittger S, Dugas M, Hiddemann W, Kern W *et al*. Molecular characterization of acute leukemias by use of microarray technology. *Genes Chromosomes Cancer* 2003; **37**: 396–405.
6 Schoch C, Kohlmann A, Schnittger S, Brors B, Dugas M, Mergenthaler S *et al*. Acute myeloid leukemias with reciprocal rearrangements can be distinguished by specific gene expression profiles. *Proc Natl Acad Sci USA* 2002; **99**: 10008–10013.
7 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**: 133–143.
8 Ferrando AA, Neuberg DS, Staunton J, Loh ML, Huard C, Raimondi SC *et al*. Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell* 2002; **1**: 75–87.
9 Dugas M, Schoch C, Schnittger S, Haferlach T, Danhauser-Riedl S, Hiddemann W *et al*. A comprehensive leukemia database: integration of cytogenetics, molecular genetics and microarray data with clinical information, cytomorphology and immunophenotyping. *Leukemia* 2001; **15**: 1805–1810.
10 Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V *et al*. NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res* 2003; **31**: 82–86.
11 Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001; **98**: 5116–5121.
12 Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998; **95**: 14863–14868.
13 Hilsenbeck SG, Friedrichs WE, Schiff R, O'Connell P, Hansen RK, Osborne CK *et al*. Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *J Natl Cancer Inst* 1999; **91**: 453–459.
14 Hubbell E, Liu WM, Mei R. Robust estimators for expression analysis. *Bioinformatics* 2002; **18**: 1585–1592.
15 Liu WM, Mei R, Di X, Ryder TB, Hubbell E, Dee S *et al*. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* 2002; **18**: 1593–1599.
16 Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS *et al*. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA* 2000; **97**: 262–267.
17 Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 2000; **16**: 906–914.
18 Vapnik V. *Statistical Learning Theory*. New York: Wiley, 1998.
19 Bene MC, Castoldi G, Knapp W, Ludwig WD, Matutes E, Orfao A *et al*. Proposals for the immunological classification of acute leukemias. European Group for the Immunological Characterization of Leukemias (EGIL). *Leukemia* 1995; **9**: 1783–1786.
20 Leo A, Wienands J, Baier G, Horejsi V, Schraven B. Adapters in lymphocyte signaling. *J Clin Invest* 2002; **109**: 301–309.
21 Rozovskaia T, Ravid-Amir O, Tillib S, Getz G, Feinstein E, Agrawal H *et al*. Expression profiles of acute lymphoblastic and myeloblastic leukemias with ALL-1 rearrangements. *Proc Natl Acad Sci USA* 2003; **100**: 7853–7858.
22 Kawagoe H, Humphries RK, Blair A, Sutherland HJ, Hogge DE. Expression of HOX genes, HOX cofactors, and MLL in phenotypically and functionally defined subpopulations of leukemic and normal human hematopoietic cells. *Leukemia* 1999; **13**: 687–698.
23 Masternak K, Muhlethaler-Mottet A, Villard J, Zufferey M, Steimle V, Reith W. CIITA is a transcriptional coactivator that is recruited to MHC class II promoters by multiple synergistic interactions with an enhanceosome complex. *Genes Dev* 2000; **14**: 1156–1166.
24 Villadangos JA, Ploegh HL. Proteolysis in MHC class II antigen presentation: who's in charge? *Immunity* 2000; **12**: 233–239.
25 Schwarz G, Brandenburg J, Reich M, Burster T, Driessen C, Kalbacher H. Characterization of legumain. *Biol Chem* 2002; **383**: 1813–1816.
26 Liu C, Sun C, Huang H, Janda K, Edgington T. Overexpression of legumain in tumors is significant for invasion/metastasis and a candidate enzymatic target for prodrug therapy. *Cancer Res* 2003; **63**: 2957–2964.
27 Weiner HL, Huang H, Zagzag D, Boyce H, Lichtenbaum R, Ziff EB. Consistent and selective expression of the discoidin domain receptor-1 tyrosine kinase in human brain tumors. *Neurosurgery* 2000; **47**: 1400–1409.
28 Dyer MJ. The role of CAMPATH-1 antibodies in the treatment of lymphoid malignancies. *Semin Oncol* 1999; **26**: 52–57.