A leukemia-enriched cDNA microarray platform identifies new transcripts with relevance to the biology of pediatric acute lymphoblastic leukemia

Cristiano De Pittà
Lucia Tombolan
Marta Campo Dell'Orto
Benedetta Accordi
Geertruy te Kronnie
Chiara Romualdi
Nicola Vitulo
Giuseppe Basso
Gerolamo Lanfranchi

Background and Objectives. Microarray gene expression profiling has been widely applied to characterize hematologic malignancies, has attributed a molecular signature to leukemia subclasses and has allowed new subclasses to be distinguished. We set out to use microarray technology to identify novel genes relevant for leukemogenesis. To this end we used a unique leukemia-enriched cDNA microarray platform.

Design and Methods. The systematic sequencing of cDNA libraries of normal and leukemic bone marrow allowed us to increase the number of genes to yield a new release of a previously generated cDNA microarray. Using this platform we analyzed the expression profiles of 4,670 genes in bone marrow samples from 18 pediatric patients with acute lymphoblastic leukemia (ALL).

Results. Expression profiling consistently distinguished the leukemia patients into three groups, those with T-ALL, B-ALL and B-ALL with *MLL/AF4* rearrangement, in agreement with the clinical classification. Our platform identified 30 genes that best discriminate these three subtypes. Using mini-array technology these 30 genes were validated in another cohort of 17 patients. In particular we identified two novel genes not previously reported: endomucin (*EMCN*) and ubiquitin specific protease 33 (USP33) that appear to be over-expressed in B-ALL relative to their expression in T-ALL.

Interpretation and Conclusions. Microarray technology not only allows the distinction between disease subclasses but also offers a chance to identify new genes involved in leukemogenesis. Our approach of using a unique platform has proven to be fruitful in identifying new genes and we suggest exploration of other malignancies using this approach.

Keywords: pediatric leukemia, microarray, cDNA libraries, gene expression profiling, mini-array.

Haematologica 2005; 90:890-898

From the CRIBI Biotechnology Centre and Dipartimento di Biologia, Università degli Studi di Padova, Padova, Italy (CDP, LT, CR, NV, GL); Laboratorio di Oncoematologia Pediatrica, Dipartimento di Pediatria, Università degli Studi di Padova, Padova, Italy (MCD'O, BA, GtK, GB).

Correspondence:
Gerolamo Lanfranchi,
CRIBI Biotechnology Center and Dipartimento di Biologia, Università degli Studi di Padova, via Ugo Bassi 58/B, 35121 Padua, Italy.
E-mail: lanfra@cribi.unipd.it

The biological behavior of cancer cells is governed, in part, by their expressed genetic repertoire, and the development of techniques to determine relative levels of RNA expression offers powerful insight into the molecular mechanisms of disease. Emerging technologies, such as DNA microarrays and chips,[1-3] facilitate gene expression analysis on an unprecedented scale, allowing the surveillance of thousands of transcripts simultaneously in a single experiment. Important clinical applications of this technology include refinement of diagnoses, prediction of prognosis, study of disease progression, and the identification of relevant gene pathways for the design of novel therapeutic strategies.[4]

Recently, the field of hematopoietic malignancies has seen an explosive increase in the applications of microarray technology for a variety of purposes.[5] Several papers have recently been published on the diagnostic application of microarrays for gene expression studies in leukemia. These include microarray-based discrimination between lymphoblastic and myeloid leukemias,[6] microarray-based identification of acute lymphoblastic leukemia with *MLL* translocations,[7-8] molecular investigations of B-cell chronic lymphoblastic leukemia (B-CLL)[9] and classification and prediction of outcome in pediatric acute lymphoblastic leukemia (ALL).[10] The majority of these papers describe a comprehensive analysis of blast cell gene expression that could provide more definite classification and stratification of patients,[11] novel prognostic tools and new insights into the pathogenesis of leukemia.[12,13] For these reasons microarrays are predicted, in the near future, to become an established routine methodology for diagnosis. Microarray platforms for transcriptome analysis typically contain thousands of oligonucleotides or cDNA, which correspond to transcripts of many different genes. The power of microarray technology lies in this large-scale type of analysis and in the possibility of studying gene expression under specific pre-defined circumstances. The method can be used to search for novel molecular changes that are associated with the development and/or prognosis of leukemia. The objective analysis of the expression level of thousands of genes leads to identification of transcripts of unknown function not previously associated with a given type of tumor.[14-15] These genes may be important in tumor cell biology including key processes of cell proliferation, adhesion, apoptosis and development of drug resistance.

The aim of our study was to identify novel candidate genes to better understand the pathogenesis and biology of childhood leukemia and to evaluate the applicability of pediatric gene expression signatures to predict ALL subtypes. For this purpose we have produced and sequenced two novel cDNA libraries from pools of normal and leukemic human bone marrow. This work has enriched the first release of our cDNA microarray platform,[16] with about 900 novel gene tags specific for studying the expression profile of childhood leukemia. In contrast to the work by Moos *et al.*,[17] in our approach the 4,670 cDNA clones were not selected by the investigators but instead obtained through systematic sequencing of special libraries in which the cDNA inserts specifically correspond to the 3'-end of mRNA.

Using this new cDNA microarray platform, named Human Array 2.0, we analyzed the gene expression profiles of 18 bone marrow samples from children affected by acute lymphoblastic leukemia in order to determine whether our platform is able to categorize some different clinical subgroups of ALL samples.

We also developed and successfully validated a diagnostic miniarray to recognize different subtypes of ALL, undertaking a blind analysis of the gene expression profile of 17 additional patients (10 B-ALL, 5 T-ALL and 2 *MLL/AF4* in a validation cohort.

Design and Methods

Construction of cDNA libraries and large-scale sequencing

Two independent cDNA libraries, BLPD and LKPD, were constructed from normal and leukemic bone marrow. A pool of patients with different subtypes of childhood leukemia were used for the LKPD library and normal human bone marrow for the BLPD library. Both libraries were produced using a strategy[18] that allows the selection of a short 3'-end region of mRNA with a very uniform size (300-600 bp), tagging each transcript with a unique probe. This feature ensures that all the cDNA clones of the collection can be amplified by polymerase chain reaction (PCR) and spotted with similar efficiency and that microarrays with uniform hybridization characteristics can be produced. More details are provided in the Supplementary Information (*s-Design and Methods*).

Microarray fabrication: Human Array 2.0 and the mini-array

The microarrays used for this project (Human Array 2.0) were constructed on glass slides by arraying PCR-amplified cDNA obtained from our archive of recombinant bacterial clones. This archive consists of 4,670 different clones obtained from systematic sequencing of 3'-end skeletal muscle, heart and bone marrow-specific cDNA libraries. Samples were spotted in duplicate on derivatized glass slides (MICROMAX Glass Slides: SuperChip™ I, PerkinElmer, Wellesley, MA, USA) using the robotic system Genpak Array 21 (Genetix, Hampshire, UK). Microarrays were UV cross-linked (Stratagene, La Jolla, CA, USA) (total power of 300 mJoules) to bind the DNA to the slides and were then processed to remove any unbound PCR product, dried and stored under vacuum at room temperature in sealed boxes.

For mini-array fabrication we used the same protocol as that described above. The only difference is the platform composition: 30 genes best discriminating B-ALL and T-ALL patients, 10 tags that are differentially expressed in all examined leukemia samples and 9 housekeeping genes used as hybridization controls. PCR products were printed in quadruplicate in each mini-array and each slide contained four mini-arrays.

Tissue samples, RNA extraction, quality control and labeling

We used excess material of bone marrow (BM) samples obtained from pediatric ALL patients[19-20] at diagnosis. Informed consent was obtained from the patients' parents. The study protocol was approved by the ethical committee of the Department of Pediatrics of the University of Padova. The patients' information relative to this study is reported in Table 1. All samples used in this work had a blast cell infiltration of more than 85%. Bone marrow mononuclear cells (BMMN) were isolated using a Lymphoprep™ density gradient (1.077 g/mL, Nycomed Pharma, Oslo, Norway) and 20-30×10^6 cells were used to prepare total RNA applying the Trizol (Invitrogen) methodology according to the manufacturers' instructions. Total RNA (100 ng) was used for quality control by capillary electrophoresis using the RNA 6000 Nano LabChip and the Agilent Bioanalyzer 2100 (Agilent

Technologies, Palo Alto, CA, USA). Only RNA samples showing absence of genomic DNA contamination and no sign of degradation were used for microarray hybridization. Two micrograms of total RNA were retro-transcribed and labeled using a MICROMAX TSA labeling kit (PerkinElmer).

Microarray hybridization

Microarray hybridization was carried out in a dual slide chamber (HybChamber, Gene Machines, San Carlos, CA, USA) humidified with 100 μL of 3×SSC. Labeled cDNA was dissolved in 40 μL of hybridization buffer, denatured at 90°C for 2 min and applied directly to the slides. Microarrays were covered with a 22×40 mm cover slip and hybridized overnight at 65°C by immersion in a high precision water bath (W28, Grant, Cambridge, UK). Post-hybridization washing was performed according to the MICROMAX TSA detection kit (PerkinElmer). Two replicates of each experiment were done using different microarray slides, in which the sample and reference RNA labeled with either Cy3 or Cy5 fluorochromes were crossed in both combinations.

Statistical analysis of expression data

Array scanning was carried out using a GSI Lumonics LITE dual confocal laser scanner with a ScanArray Microarray Analysis System (Perkin Elmer), and raw images were analyzed with QuantArray Analysis Software (GSI Lumonics, Ottawa, Canada). Normalization of expression levels[21] of all spot replicates was performed with MIDAS (TIGR Microarray Data Analysis System). We applied the Lowess (Locfit) normalization to expression data before any other statistical analysis. After normalization, values of spot replicates in all array experiments were averaged. Identification of differentially expressed genes was performed with Significance Analysis of Microarray (SAM).[22] In addition, a set of discriminant genes (putative markers) was selected using Predictive Analysis of Microarray (PAM).[23] Principal component analysis, cluster analysis, k-means and profile similarity searching were performed with R software[24] and J-Express.[25]

Table 1. **Patients' description.**

Patient	*Phenotype*	*Molecular translocations: MLL/AF4; Bcr/Abl; TEL/AML1; E2A/PBX1*	*Age at diagnosis*	*Protocol*
1A				
1	CALL	negative	1y6m	95
2	CALL	negative	4y1m	95
3	CALL	negative	12y9m	95
4	CALL	negative	5y5m	95
5	CALL	negative	13y5m	LAL 2000
6	CALL	*TEL/AML1*	2y10m	LAL 2000
7	CALL	*TEL/AML1*	8y7m	LAL 2000
8	Pre-B	*E2A/PBX1*	14y3m	LAL 2000
9	CALL	negative	9y11m	LAL 2000
10	CALL	negative	8y5m	LAL 2000
11	T-ALL	negative	12y10m	LAL 2000
12	T-ALL	negative	5y1m	LAL 2000
13	T-ALL	negative	9y4m	LAL 2000
14	T-ALL	negative	4y	LAL 2000
15	T-ALL	negative	12y4m	LAL 2000
16	Pre-Pre-B	*MLL/AF4*	4 y10m	LAL 2000
17	Pre-Pre-B	*MLL/AF4*	14y3m	LAL 2000
18	Pre-B	*MLL/AF4*	4y10m	95
Patient	*Phenotype*	*Molecular translocations: MLL/AF4; Bcr/Abl; TEL/AML1; E2A/PBX1*	*Age at diagnosis*	*Protocol*
1B				
19	CALL	TEL/AML1	2y1m	LAL2000
20	CALL	negative	5y1m	LAL2000
21	CALL	negative	2y	LAL2000
22	CALL	negative	333y11m	LAL 2000
23	CALL	negative	13y6m	LAL 2000
24	CALL	negative	5y7m	LAL 2000
25	PreB	negative	1y11m	LAL2000
26	PreB	negative	10y4m	LAL 2000
27	PreB	negative	4y9m	LAL 2000
28	PreB	negative	8y9m	LAL 2000
29	T-ALL	negative	4y9m	LAL 2000
30	T-ALL	negative	7y6m	LAL 2000
31	T-ALL	negative	3y	LAL 2000
32	T-ALL	negative	5y	LAL 2000
33	T-ALL	negative	9y3m	LAL 2000
34	PreB	*MLL/AF4*	4y10m	LAL 2000
35	Pre-Pre-B	*MLL/AF4*	12y3m	LAL 2000

Phenotypic characterization of patients was performed by flow cytometry using a direct immunofluorescence technique with four-color combinations of monoclonal antibodies. The panel of markers is routinely employed for diagnosis of acute leukemias at the AIEOP reference laboratory for immunophenotypic studies of the University of Padova. Specific chromosomal aberrations MLL/AF4, Bcr/Abl, TEL/AML1 and E2A/PBX1 were identified through molecular studies (i.e. RT-PCR assays) according to the Biomed-1 protocol. Pre-B: pre-B-ALL, pre-pre-B: pre-pre-B-ALL; CALL: common-ALL; T-ALL: T-lineage ALL) (see reference 42 for phenotypic definitions of childhood ALL). ***A****. patients' samples studied with the Human Array 2.0;* ***B****. patients' samples studied with the mini-array.*

Results

Differential profiling of normal and leukemic bone marrow by counting expressed sequence tags

The 3'-end specific cDNA clones of two independent cDNA libraries, named BLPD (normal bone marrow) and LKPD (leukemic bone marrow), have been sequenced (for more information see s-Design and methods). In total we have produced about 2,000 expressed sequence tags (EST) for the BLPD library and about 4,000 EST for the LKPD libraries. Of these, 1,535 EST from the BLPD library and 2,983 EST from the LKPD library that passed our quality standards were assembled into similarity groups. The EST described in this paper have been deposited in the EBI-GenBank-DBJ public database (accession numbers for BLPD *AJ705105-AJ706640* and for LKPD *AJ712340-AJ715323*). For each EST group, a consensus sequence was produced and searched on public databases. Finally, we ended up with an EST collection that identifies 782 transcripts of normal bone marrow and 1,298 transcripts of leukemic bone marrow. After completion of the first sets of 1,920 EST of both libraries, we realized that about 40% of the EST of the BLPD library corresponded to 6 very abundant transcripts (Supplemental Data, s-Figure 1). The presence of highly represented tissue-specific mRNA is a common feature of strictly committed organs, such as skeletal muscle, heart, liver and pancreas. Their presence greatly hampers the identification of rare transcripts by a random sequencing approach. We also experienced a similar situation in skeletal muscle, for which five mRNA of mitochondrial origin and three mRNA transcribed from the nuclear DNA accounted for 42% of the total skeletal muscle mRNA population.[26,27] For this reason we decided to continue sequencing only the pathological bone marrow cDNA library, in which the percentage of EST that identifies abundant transcripts is reduced to 10%, while more than 50% recognize low abundance transcripts.

The systematic sequencing of EST from unbiased cDNA libraries is a suitable approach for analyzing the gene expression profile of a given tissue.[28] In fact, the frequency of a given EST in the cDNA library can be related to the relative abundance of the corresponding mRNA in the source tissue. In addition, the use of short tags specifically taken from the 3'-end of mRNA has the further advantages of providing a unique identification of a given transcript and possible discrimination of gene isoforms, since the 3' end of mRNA is the most divergent part of a gene. We performed a statistical analysis to define the transcriptional profile in the normal and pathological bone marrow obtained by EST counting. Considering 1,920 EST of both libraries independently, we clustered them by identity and compared the number of EST sequences belonging to each pair of clusters of the two libraries that identify the same transcript. In this way we found about 30 genes that appear to be differentially expressed in the leukemic bone marrow (Supplemental Data, s-Table 1). We used three different statistical tests to check this data, and obtained analogous results. In particular, we found that a series of ribosomal genes are over-expressed in pathological bone marrow. A similar situation was found in colon carcinoma for which the mRNA level of S19 mRNA was higher in primary colon carcinoma tissue than in matched normal colon tissue.[29] In contrast, S100 calcium binding proteins (*S100A8* and *S100A9*), hemoglobin α 1 (HBA1) and defensin alpha 1 (*DEFA1*) were found to be strongly down-regulated in the leukemic cDNA library. We exploited quantitative reverse transcriptase-PCR, using the SYBR-Green method,[30] to quantify the level of expression of some transcripts in leukemic bone marrow mRNA, in order to validate the results obtained from the statistical analysis. We selected four genes, *RPL13A*, *TPT1*, *RPL37A* and *RPL10A*, that have levels of expression representative of the entire scale of variation of gene expression. The housekeeping gene glyceraldehyde-3-phosphate dehydrogenase (*GADPH*) was used as the control. The expression values obtained with the quantitative reverse transcriptase-PCR for all tested transcripts were in agreement with the statistical comparison of EST frequencies in the two libraries.

The Human Array 2.0

The EST counting approach is not appropriate for expression profiling of individual cases of childhood leukemia because it is impossible to build and sequence a specific 3'-end cDNA library for each patient. Therefore, for this further study we decided to use microarray technology. The results of the systematic sequencing described in the previous paragraph (782 and 1,298 cDNA clones identifying unique

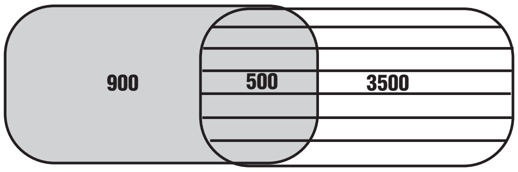

Figure 1. Probe composition of Human Array 2.0. This cDNA platform contains a large number of clones (70%) suitable for studying the muscle transcriptome, but also a portion of genes (20%) relevant for expression profiling of tumors, in particular childhood leukemia. Five hundred probes, probably housekeeping genes expressed in a wide spectrum of tissues, are in common between the muscle and leukemia components of our platform.

□ Transcripts identified by systematic sequencing of normal and leukemic bone marrow cDNA libraries (BLPD and LKPD)
▤ Transcripts identified in all cDNA libraries
▤ Transcripts identified by systematic sequencing of skeletal muscle cDNA libraries (HSPD, CMPD)

transcripts from normal and leukemic bone marrows, respectively) were used to enrich a cDNA microarray platform that was already available and established in our laboratory. This platform derives mainly from muscles (skeletal and heart) but contains probes for many housekeeping genes expressed in a wide spectrum of tissues, which justify its use in more general profiling studies. Nevertheless, to complete the platform and to better address this specific research, we compared the list of available probes with the collection of unique transcript tags obtained from the normal and leukemic bone marrow. The result of this comparison is reported in Figure 1: 3,257 probes come specifically from skeletal and cardiac muscles, 500 clones are in common and 913 clones (20%) represent the specific contribution of the bone marrow libraries. These novel probes were therefore added to the original platform to obtain the Human Array 2.0.

Expression profiling of single patients with childhood leukemia

To determine the gene expression profiling of single leukemic samples, we analyzed 18 diagnosed bone marrow samples with the Human Array 2.0 containing 4,670 cDNA probes (GEO Platform No. GPL2011). The clinical characteristics of these patients are detailed in Table 1A. The 18 samples came from children with acute lymphoblastic leukemia (ALL). These ALL patients were subdivided into three different subtypes by morphological, immunophenotypic and cytogenetic parameters: 5 had T-ALL, 10 had B-ALL and 3 had B-ALL with the specific chromosomal aberration *MLL/AF4*. We compared the transcription profiles of the leukemic samples by microarray competitive hybridization against an arbitrary total RNA reference prepared from normal human bone marrow (BD Biosciences, San Josè, CA, USA). Initially, as a general analysis, we used a two-dimensional hierarchical clustering algorithm showing that the expression profiling divided the 18 patients into three distinct groups that corresponded perfectly to the clinical classification (Figure 2). This result was confirmed by applying different distance measures (Euclidean and Pearson correlation) to our data. Figure 2 clearly shows that our microarray experiments are able to categorize different clinical subgroups of ALL samples, namely B-ALL, T-ALL and *MLL/AF4* (GEO Series No. GSE2604). Moreover, the expression profiles reported here show that ALL samples possessing a rearranged *MLL* (3 patients) have a highly uniform and distinct pattern that distinguishes them from conventional ALL. As reported by Armstrong *et al.*,[31] MLL is characterized by the expression of myeloid-specific genes, and this raises the possibility that *MLL* may be more closely related to AML. Only for patient n. 10 did the diagnosis and molecular classification not agree. The hierarchical clustering showed that the expression profile of patient n. 10 is similar to that of patients with a *MLL/AF4* rearrangement (Figure 2). According to diagnostic criteria this child was affected by B-ALL and was reconfirmed not to have the *MLL/AF4* rearrangement.

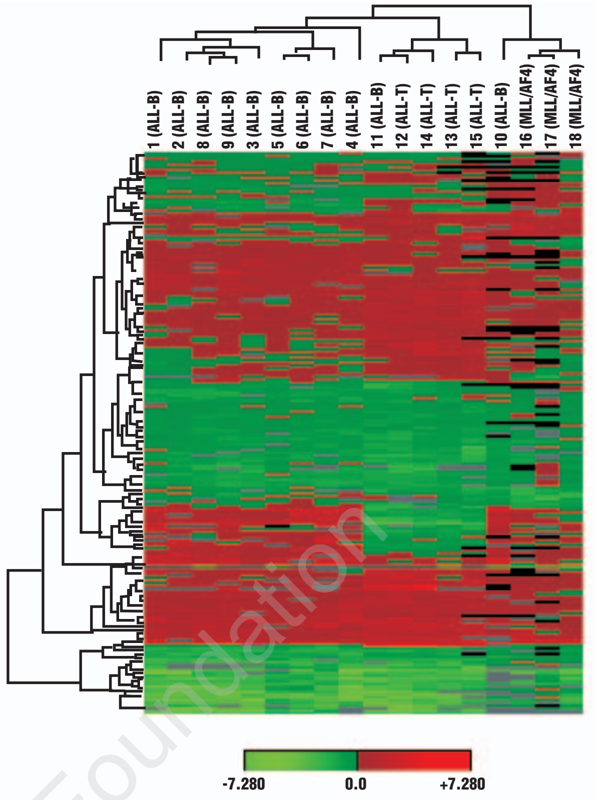

Figure 2. Hierarchical cluster analysis with Euclidean distance and complete linkage method on gene expression profiles of 18 patients with childhood leukemia (see Table 1A) with different clinical traits: B-ALL (n=10), T-ALL (n=5) and *MLL/AF4* (n=3). This analysis was based on about 200 differentially expressed genes, which were identified using the SAM software package in a supervised approach, considering the three entities B-ALL, T-ALL and *MLL/AF4*. Patients are clearly divided into three groups and this shows that the molecular analysis is in agreement with the clinical diagnosis. Expression values of these transcripts are provided in the Supplementary Information (s-Table 2). The dendrogram was created with J-Express software. A color-coded scale for the deviation compared to a z-score based mean has been used: red cells represent high expression and green cells low expression levels in comparison to normal human bone-marrow.

Differentially expressed transcripts common to all 18 patients

Analyzing the expression profiles of patients using the SAM program, we were able to identify, with a 13% false discovery rate, about 200 common differentially expressed genes (Supplemental Data, s-Table 2). Forty-six percent of these deregulated transcripts were detected in the microarray platform by probes that had been obtained by the systematic sequencing of normal and leukemic bone marrow libraries.

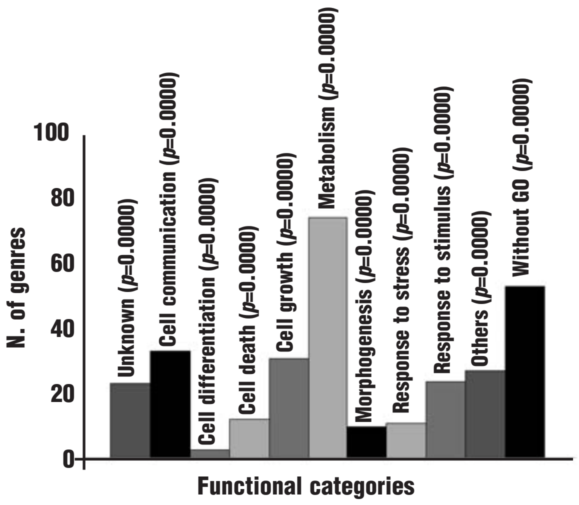

Figure 3. The transcripts that were differentially expressed in all 18 patients are grouped here according to their biological function in ten different classes. This functional classification was made following the criteria of the FATIGO program (*http://fatigo.bioinfo.cnio.es*). The *p* value of statistical significance for each category, calculated by the R statistical software, is reported. (GO: gene ontology).

Therefore, these cDNA clones, which account for only 20% of the total probe set in the platform, provide a fundamental contribution to the correct classification of patients affected by childhood ALL. These differentially expressed genes have been classified according to their function with FATIGO (available at URL *http://fatigo.bioinfo.cnio.es*) to be associated with ten different biological processes (Figure 3). Many of them appear to be involved in processes such as cell communication, differentiation, cell death and growth and therefore could play an important role in leukemia biology.

In this context, three genes showed a significant upregulation: stathmin 1 (*STMN1*), phosphodiesterase 7A (*PDE7A*) and prostate tumor over-expressed gene 1 (*PTOV1*) while one gene, TYRO protein tyrosine kinase binding protein (*TYROBP* also called *DAP12*), was generally under-expressed. *STMN1* encodes a ubiquitous cytosolic phosphoprotein supposed to function as an intracellular transducer integrating regulatory signals of the cellular environment. This protein is induced in normal lymphocytes following mitogenic stimulation and is involved in the destabilization of microtubule filaments.[32] Recently *STMN1* was found to be over-expressed in a subgroup of human breast carcinomas where it was supposed to play a role in the control of tumor cell growth and differentiation.[33] The expression of *PDE7A1* has been detected in T cells and hence we found this transcript particularly over-expressed in patients with T-ALL. This phosphodiesterase plays a role in signal transduction by regulating the intracellular concentration of cyclic nucleotides, being highly specific for cAMP.[34] The PTOV1 protein consists of two novel protein domains arranged in tandem, without significant similarities to known protein motifs. This gene is over-expressed in a significant proportion of prostate cancers and could have a role in the biology of these tumors.[35] This idea has been strengthened by the work of Santamaria *et al*.;[36] these authors have associated high in vivo levels of PTOV1 mRNA and nuclear localization of the protein with the proliferative state of carcinoma cells, while nuclear exclusion of PTOV1 *in vitro* was associated with cell quiescence. Over-expression of *PTOV1* results in nuclear localization of the protein and consequent induction of proliferation and entry into the S phase of the cell cycle. We observe here for the first time, the upregulation of this gene in ALL tumors, thus providing additional evidence that confirms the role of *PTOV1* in the control of neoplastic cell proliferation. DAP12 is a transmembrane adapter well known for its role in transducing activation signals for an extended group of receptors in natural killer (NK) cells, granulocytes, monocytes/macrophages, and dendritic cells. This gene is expressed in human NK cell lines, but not in T-leukemia or B-lymphoblastic cell lines. Lanier *et al*.[37] suggested that DAP12 protein associates with the non-inhibitory isoforms of the KIR molecules in NK cells and promotes cellular activation through these receptors, in a way similar to the function of the CD3 subunits in the T-cell receptor complex or the CD79 subunits in the B-cell receptor complex. Moreover, Aoki *et al*. reported that signaling through DAP12 was very important for terminal differentiation of the murine M1 leukemia cell line;[38] they observed a morphological change of M1 cells to macrophages, including giant cell formation after stimulation through DAP12. Accordingly, in our experiments *DAP12* appeared to be down-regulated in all ALL. We found a large group of deregulated genes belonging to the protein biosynthetic class that includes nearly all the ribosomal proteins; they appeared generally over-expressed. This confirms our data obtained by the systematic sequencing of the leukemic libraries and by quantitative real time PCR tests.

Figure 3 shows that many differentially expressed transcripts have as yet unknown functions. In this group we found 22 genes that exhibited very significant levels of over- or under-expression. In the future it will be interesting to investigate the role of some of these unknown genes in the pathogenesis of leukemia since they could be new candidate genes for better understanding the behavior of these tumors.

Transcripts discriminating ALL subtypes

Using PAM algorithms we identified approximately 30 deregulated transcripts that differentiate the three subtypes of ALL well (Supplemental Data, s-Table 3). In particular, we found genes that distinguished B-lineage ALL *versus* T-lineage ALL. Furthermore, several biologically predictable patterns of differential gene expression were observed. Class II HLA molecules are located predominantly on B cells and macrophages, playing a major role in antigen presentation. Consistent with known endogenous patterns of expression, MHC molecules were over-expressed in B-lineage ALL when compared with T-ALL. Of the top 30 genes discriminating T-lineage versus B-lineage ALL, four were class II MHC molecules (*HLA-DPA1*, *HLA-DQB1*, *HLA-DRA*, *HLA-DRB1*). Similarly, the delta subunit of the T-cell antigen receptor CD3D was highly expressed in T-ALL samples. We identified specific molecular markers for T or B lineages which are already used in clinical diagnosis, but in addition we discovered discriminating genes that might become new molecular markers for better classification of patients.

As envisaged earlier,[39] the genes required to diagnose all known ALL subtypes may be only several dozen. A panel of selected genes may thus form the basis of a diagnostic tool. Therefore, we built a diagnostic mini-array composed of the 30 discriminating genes determined by PAM analysis of the expression datasets obtained with the Human Array 2.0 platform. We tested this mini-array on a novel cohort of patients for its discriminating capacity. The results of these experiments are reported in Figure 4. As can be seen, the mini-array allows a clear separation of T-ALL and B-ALL. Moreover, two patients with a *MLL/AF4* aberration clustered together as a subgroup of ALL-B, indicating that these two samples are most similar. These further experiments confirmed the analyses of the expression profiles of childhood leukemia patients obtained with the more dense microarray platform and show that the mini-array might be a good basic diagnostic tool for childhood leukemia that can be improved with further marker probes to increase its potential resolution.

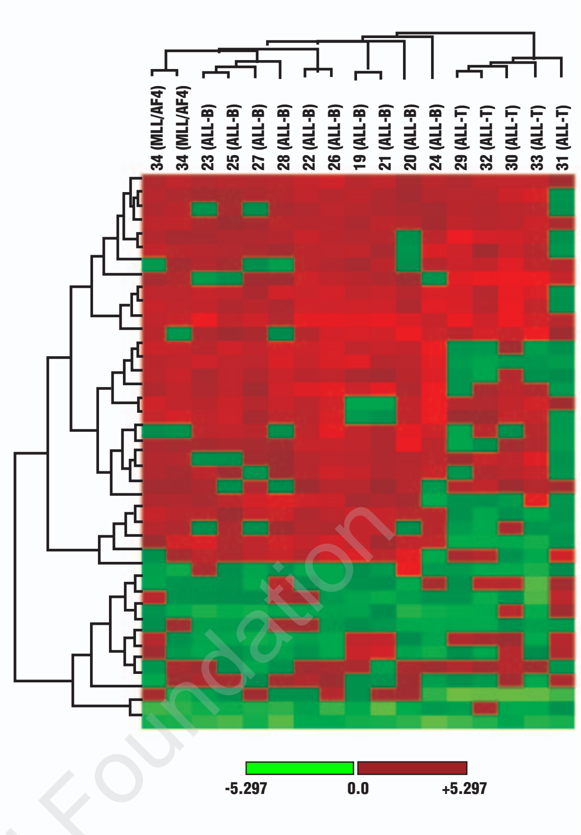

Figure 4. Hierarchical cluster analysis with Euclidean Distance and complete linkage method on gene expression profiles of 17 childhood patients (see Table 1B) with different clinical characteristics: T-ALL (n=5), B-ALL (n=10) and *MLL/AF4* (n=2). We obtained these results using the diagnostic mini-array composed of 40 different gene probes. Patients are divided into three groups in agreement with the clinical diagnosis: B-ALL samples are clearly separate from T-ALL and *MLL/AF4* patients are a subgroup of B-ALL. Each column represents a profile of an ALL sample, and each row contains values referring to an individual gene. We used the color-coded scale for the normalized expression values as in Figure 3. Expression values of each transcript are provided in the Supplementary Information (s-Table 4).

Comparison with other leukemia expression datasets

We compared the discriminating genes found with our microarray analysis with the genes differentiating MLL versus ALL and AML reported by Armstrong *et al*. after oligonucleotide chip analysis.[31] Many of them are in common, but seven of our 30 discriminating genes are not present in the Affymetrix U95A oligonucleotide arrays used by these authors, namely *SIGIRR*, *PDE7A*, *GSPT2*, *USP33*, *BM045*, *MORF4L1* and *EMCN*. In particular, *USP33* and *EMCN* were over expressed in all B-ALL and *MLL-AF4* patients relative to T-ALL cases. USP33 also called VDU1 (VHL-interacting deubiquitinating enzyme 1) was identified in a two-hybrid screen to interact directly with pVHL, the von Hippel-Lindau protein. VDU1 belongs to the USP family of deubiquitinating enzymes that compete or counteract with the ubiquitin-conjugating system by rescuing proteins from degradation by removing ubiquitin from protein substrates.[40] The differential expression of USP33 in ALL is remarkable and strongly merits the study of USP33 function in normal lymphocytes and leukemias. The human endomucin gene (EMCN) encodes a protein with a transmembrane domain and multiple glycosylation sites. As a membrane protein, human endomucin may have a role in one or more steps of angiogenesis such as endothelial proliferation, migration or differentiation into tube-like structures. It also mediates the interaction between endothelial cells with other cell types, such as those of the hematopoietic lineage during inflammation, or with cancer cells during metastasis.[41] It is still unclear at present whether endomucin has adhesive or anti-adhesive activity or whether it might be involved in signal transduction.

Discussion

Our study on gene expression in childhood leukemia was structured in two parts. In the first set of experiments we produced 3'-end specific cDNA libraries from normal and leukemic bone marrow. Systematic sequencing of these libraries allowed us to increase the number of useful genes to yield a new release of our cDNA microarray and to have a general picture of the differences in gene expression at least in the range of abundantly transcribed mRNA. Our cDNA platform identifies a large number of clones (70%) suitable for studying the muscle transcriptome and neuromuscular diseases, but also a portion of genes (20%) relevant for gene expression profiling of childhood leukemia. In a second set of experiments we analyzed 18 patients affected by childhood leukemia with the Human Array 2.0 containing 4,670 probes. These RNA samples were subdivided into 3 different clinical classes: 5 T-ALL, 10 B-ALL and 3 *MLL/AF4*. We have demonstrated that our microarray platform is able to identify 3 different groups in agreement with traditional diagnostic measures such as clinical features, immunophenotype and cytogenetic alterations. This result is important because although only 20% of gene probes in our platform are specific for bone marrow, the platform nevertheless appears to be adequate for distinguishing different clinical subtypes of childhood leukemia. Microarray methodology has been used successfully to study many hematopoietic and solid tumors. Golub *et al*. have shown that array technology can reliably classify leukemia and distinguish acute myeloid from lymphoid leukemia as well as B- from T-lineage ALL.[6] More recently, a large number of leukemia patients were studied, demonstrating that expression profiling can distinguish T-lineage ALL and all the common B-lineage chromosomal translocations.[10] In these studies the bone marrow samples were analyzed with Affymetrix chips containing about 12,600 oligonucleotide probes. In our work we obtained the same results with a cDNA platform of only 4,670 genes. We were able to correctly classify patients with this reduced platform because of the presence of some 900 cDNA probes specifically derived by the sequencing of leukemic bone marrow libraries. In fact, 46% of differentially expressed genes identified by the SAM analysis belong to the bone marrow component of our microarray platform. Furthermore the use of this platform has revealed a novel group of transcripts commonly altered in ALL patients (*STMN1*, *PDE7A*, *PTOV1* and *TYROBP*) which could be important in the biology of childhood leukemia. We found that no single transcript could unilaterally distinguish subgroups, but instead we identified about 30 deregulated genes that consistently discriminated among three subtypes of ALL, in particular B-lineage versus T-lineage. Using mini-array technology these 30 genes were validated on a new cohort of 17 patients (GEO series No. GSE2617). Our diagnostic mini-array confirmed the distinction between different subtypes of ALL. We propose that this minimal microarray platform could have many advantages as a molecular diagnostic tool, including high sensitivity, high specificity, and straightforward interpretation of the expression data.

CDP: construction of the cDNA libraries, systematic sequencing of the cDNA libraries, construction and control of the microarray platform, microararry experiments, microarray data analysis; LT: construction of the cDNA libraries, systematic sequencing of the cDNA libraries, construction and control of the microarray platform, RNA sample preparation and control, microarray data analysis; GtK: interpretation and discussion of microarray expression data, revision of the manuscript; CR: management and statistical analysis of microarray datasets; NV: bioinformatic analysis of cDNA libraries sequence data, clustering of EST, annotation of EST; GB: project planning, leukemia case identification and clinical management, human samples; GL: project planning, concept and design, data organization and interpretation, manuscript writing. The authors declare that they have no potential conflicts of interests. We thank Rosanna Zimbello and Barbara Simionati of the CRIBI Sequencing Service (URL: http://bmr.cribi.unipd.it) for help in cDNA sequencing. We are also grateful to Beniamina Pacchioni and Stefano Cagnin for their help with microarray fabrication and hybridization.

The work described in this paper was financed by the Fondazione Città della Speranza ONLUS, Padova, Italy, by CNR and Ministero dell'Università e della Ricerca (Cofin, FIRB and 60%). The instrumentation for microarray construction and analysis for this study was purchased thanks to a generous donation from the Fondazione della Cassa di Risparmio di Padova e Rovigo, Padova, Italy.

On-line supplementary information is available at URL http://muscle.cribi.unipd.it/microarrays/leukemia/. It contains complete sets of expression data and where lists or profiles of altered transcripts found in our experiments. MIAME standards have been followed in microarray data analysis and presentation. Expression datasets have been submitted to the GEO database (http://www.ncbi.nlm.nih.gov/geo/). The accession number for the Human Array 2.0 platform is GPL2011 and those for the expression datasets are GSE2604 and GSE2617 series.

Manuscript received August 10, 2004. Accepted May 15, 2005.

References

1. Schena M, Shalon D, Davies RW and Brown PO. Quantitative monitoring of gene expression patterns with complementary DNA microarray. Science 1995; 270:467-70.
2. Schena M, Shalon D, Heller R, Chai A, Browns PO, Davis RW. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci USA 1996;93:10614-9.
3. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Mittmann M, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996; 14:1675-80.
4. Young RA. Biomedical discovery with DNA arrays. Cell 2000;102:9-15.
5. Levene AP, Morgan GJ, Davies FE. The use of genetic microarray analysis to classify and predict prognosis in haematological malignancies. Clin Lab Haematol 2003;25:209-20.
6. 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: 531-7.
7. Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet 2002;30:41-7.
8. Tsutsumi S, Taketani T, Nishimura K, Ge X, Taki T, Sugita K, et al. Two distinct gene expression signatures in pediatric acute lymphoblastic leukemia with MLL rearrangements. Cancer Res 2003; 63:4882-7.
9. Rosenwald A, Alizadeh AA, Widhopf G, Simon R, Davis RE, Yu X. Relation of gene expression phenotype to immunoglobulin mutation genotype in B cell chronic lymphocytic leukemia. J Exp Med 2001; 194:1639-47.
10. 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; 1:133-43.
11. Kohlmann A, Schoch C, Schnittger S, Dugas M, Hiddemann W, Kern W, et al. Pediatric acute lymphoblastic leukemia (ALL) gene expression signatures classify an independent cohort of adult ALL patients. Leukemia 2004;18:63-71.
12. Haferlach T, Kohlmann A, Kern W, Hiddemann W, Schnittger S, Schoch C. Gene expression profiling as a tool for the diagnosis of acute leukemias. Semin Hematol 2003;40:281-95.
13. Larramendy ML, Niini T, Elonen E, Nagy B, Ollila J, Vihinen M, Knuutila S. Overexpression of translocation-associated fusion genes of FGFRI, MYC, NPMI, and DEK, but absence of the translocations in acute myeloid leukemia. A microarray analysis. Haematologica 2002; 87:569-77.
14. Niini T, Vettenranta K, Hollmen J, Larramendy MI, Aalto Y, Wikman H, et al. Leukemia 2002;16:2213-21.
15. Staal FJT, van der Burg M, Wessels LFA, Barendregt BH, Baert MRM, van der Burg CMM, et al. DNA microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-B acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers. Leukemia 2003;17: 1324-32.
16. Campanaro S, Romualdi C, Fanin M, Celegato B, Pacchioni B, Trevisan S, et al. Gene expression profiling in dysferlinopathies using a dedicated muscle microarray. Hum Mol Genet 2002;11: 3283-98.
17. Moos PJ, Raetz EA, Carlson MA, Szabo A, Smith FE, Willman C, et al. Identification of gene expression profiles that segregate patients with childhood leukemia. Clin Cancer Res 2002; 8:3118-30.
18. Lanfranchi G, Muraro T, Caldara F, Pacchioni B, Pallavicini A, Pandolfo D, et al. Identification of 4370 expressed sequence tags from a 3'-end-specific cDNA library of human skeletal muscle by DNA sequencing and filter hybridization. Genome Res 1996;6:35-42.
19. Basso G, Buldini B, De Zen L, Orfao A. New methodologic approaches for immunophenotyping acute leukemias. Haematologica 2001;86:675-92.
20. De Zen L, Bicciato S, te Kronnie G, Basso G. Computational analysis of flow-cytometry antigen expression profiles in childhood acute lymphoblastic leukemia: an MLL/AF4 identification. Leukemia 2003;17:1557-65.
21. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, et al. TM4: a free, open-source system for microarray data management and analysis. Biotechniques 2003;34:374-8.
22. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci 2001;98: 5116-21.
23. 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; 99:6567-72.
24. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: a software development project. Genome Biol 2004;5:R80.
25. Dysvik B, Jonassen I. J-Express: exploring gene expression data using Java. Bioinformatics 2001;17:369-70.
26. Pacchioni B, Trevisan S, Gomirato S, Toppo S, Valle G, Lanfranchi G. Semi-multiplex PCR technique for screening of abundant transcripts during systematic sequencing of cDNA libraries. Biotechniques 1996;21:644-9.
27. Laveder P, De Pitta C, Toppo S, Valle G, Lanfranchi G. A two-step strategy for constructing specifically self-subtracted cDNA libraries. Nucleic Acids Res 2002; 30:e38.
28. Bortoluzzi S, Danieli GA. Towards an in silico analysis of transcription patterns. Trends Genet 1999;15:118-9.
29. Kondoh N, Schweinfest CW, Henderson KW, Papas TS. Differential expression of S19 ribosomal protein, laminin-binding protein, and human lymphocyte antigen class I messenger RNAs associated with colon carcinoma progression and differentiation. Cancer Res 1992;52:791-6.
30. Rajeevan MS, Vernon SD, Taysavang N, Unger ER. Validation of array-based gene expression profiles by real time (kinetic) RT-PCR. J Mol Diagn 2001;3: 26-31.
31. Armstrong SA, Staunton JE, Silverman LB, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet 2002;30:41-7.
32. Cassimeris L. The oncoprotein 18/stathmin family of microtubule destabilizers. Curr Opin Cell Biol 2002; 14:18-24.
33. Curmi PA, Nogues C, Lachkar S, Carelle N, Gonthier MP, Sobel A, et al. Overexpression of stathmin in breast carcinomas points out to highly proliferative tumors. Br J Cancer 2000;82: 142-50.
34. Lee R, Wolda S, Moon E, Esselstyn J, Hertel C, Lerner A. PDE7A is expressed in human B-lymphocytes and is up-regulated by elevation of intracellular cAMP. Cell Signal 2002;14:277-84.
35. Benedit P, Paciucci R, Thomson TM, Valeri M, Nadal M, Caceres C, et al. PTOV1, a novel protein overexpressed in prostate cancer containing a new class of protein homology blocks. Oncogene 2001;20:1455-64.
36. Santamaria A, Fernandez PL, Farré X, Benedit P, Reventos J, Morote J, et al. PTOV1, a novel protein overexpressed in prostate cancer, shuttles between the cytoplasm and the nucleus and promotes entry into the S phase of the cell division cycle. Am J Pathol 2003;162: 897-905.
37. Lanier LL, Corliss BC, Wo J, Leong C, Phillips JH. Immunoreceptor DAP12 bearing a tyrosine-based activation motif is involved in activating NK cells. Nature 1998;391:703-7.
38. Aoki N, Kimura S, Oikawa K, Nochi H, Atsuta Y, Kobayashi H, et al. DAP12 ITAM motif regulates differentiation and apoptosis in M1 leukemia cell line. Biochem Biophys Res Commun 2002; 291:296-304.
39. te Kronnie G, Bicciato S, Basso G. Acute leukemia subclassification: a marker protein expression perspective. Hematology 2004;9:165-70.
40. Li Z, Na X, Wang D, Schoen SR, Messing EM, Wu G. Ubiquitination of a novel deubiquitinating enzyme requires direct binding to von Hippel-Lindau tumor suppressor protein. J Biol Chem 2002;277:4656-62.
41. Liu C, Shao ZM, Zhang L, Beatty P, Sartippour M, Lane T, et al. Human endomucin is an endothelial marker. Biochem Biophys Res Commun 2001; 288:129-36.
42. van der Does-van den Berg A, Bartram CR, Basso G, Benoit YC, Biondi A, Debatin KM, et al. Minimal requirements for the diagnosis, classification, and evaluation of the treatment of childhood acute lymphoblastic leukemia (ALL) in the "BFM Family" Cooperative Group. Med Pediatr Oncol 1992;20:497-505.