Leukemia (2003) 17, 1324–1332

www.nature.com/leu

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

FJT Staal[1], M van der Burg[1], LFA Wessels[2], BH Barendregt[1], MRM Baert[1], CMM van den Burg[1], C Van Huffel[1], AW Langerak[1], VHJ van der Velden[1], MJT Reinders[2] and JJM van Dongen[1]

[1]*Department of Immunology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands; and* [2]*Department of Mediametics, Technical University Delft, Delft, The Netherlands*

Microarrays for gene expression profiling are rapidly becoming important research tools for the identification of novel markers, for example, for novel classification of leukemias and lymphomas. Here, we review the considerations and infrastructure for microarray experiments. These considerations are illustrated via a microarray-based comparison of gene expression profiles of paired diagnosis–relapse samples from patients with precursor-B acute lymphoblastic leukemia (ALL), who relapsed during therapy or after completion of treatment. Initial experiments showed that several seemingly differentially expressed genes were actually derived from contaminating non-leukemic cells, particularly myeloid cells and T-lymphocytes. Therefore, we purified the ALL cells of the diagnosis and relapse samples if their frequency was lower than 95%. Furthermore, we observed in earlier studies that extra RNA amplification leads to skewing of particular gene transcripts. Sufficient (non-amplified) RNA of purified and paired diagnosis–relapse samples was obtained from only seven cases. The gene expression profiles were evaluated with Affymetrix U95A chips containing 12 600 human genes. These diagnosis–relapse comparisons revealed only a small number of genes ($n=6$) that differed significantly in expression: mostly signaling molecules and transcription factors involved in cell proliferation and cell survival were highly upregulated at relapse, but we did not observe any increase in drug-resistance markers. This finding fits with the observation that tumors with a high proliferation index have a poor prognosis. The genes that changed between diagnosis and relapse are currently not in use as diagnostic or disease progression markers, but represent potential new markers for such applications.
Leukemia (2003) **17,** 1324–1332. doi:10.1038/sj.leu.2402974
Keywords: gene expression profiling; DNA microarray; precursor-B-ALL; relapse

Introduction

DNA microarrays for the investigation of gene expression profiles may very well revolutionize basic biological research and laboratory investigations of patient material. Especially in the field of hematopoietic malignancies, application of microarray technology has already been used for studying patient series. Several seminal papers on the diagnostic application of microarrays for gene expression studies in leukemia and lymphoma have recently been published.[1–6] These include a microarray-based discrimination between lymphoblastic and myeloid leukemias,[5] a clinically relevant subdivision of diffuse large B-cell lymphoma,[1] microarray-based identification of *MLL* gene aberrations in acute lymphoblastic leukemia (ALL)[2] and investigations in B-CLL.[6]

Correspondence: Professor Dr JJM van Dongen, Department of Immunology, Erasmus MC, University Medical Center Rotterdam, Dr Molewaterplein 50, Rotterdam 3015 GE, The Netherlands; Fax: +31 10 4089 456
Received 15 November 2002; accepted 28 February 2003

Gene expression microarrays typically contain thousands of oligonucleotides or cDNAs, which correspond to transcripts of many different genes. The power of microarrays lies in the high numbers and in the possibility to study expression of genes under specific (pre-defined) circumstances. The objective analysis of the expression level of thousands of genes almost automatically leads to identification of previously unknown functions for several genes.

DNA microarrays come in two main types of technical platforms. The first is based on standard microscopic glass slides on which cDNAs or long oligonucleotides (typically 70–80 mers) have been spotted. The second is based on photolithographic techniques to synthesize 25-mer oligonucleotides on a silicon wafer and constitutes the patented technology of Affymetrix Inc. Consequently, the production methods of the two types of DNA microarrays are different: glass slide arrays can be produced relatively easily (if the appropriate spotting equipment is readily available), but the Affymetrix microarrays need highly specialized production facilities. The two platforms also differ in their technology for labeling of RNA molecules. However, the main difference concerns the method to assess the gene transcript levels: quantitation via pairwise comparisons (ratios) for glass slides or quantitations in arbitrary (but well-defined) expression units in the case of Affymetrix arrays. The differences between the two main types of platforms result in specific advantages and disadvantages, which need to be weighed carefully per type of application before the appropriate choice can be made.

Glass slide microarrays for gene expression

In glass slide microarray studies, RNA from the target sample and from the 'control' sample are pairwise studied as an equivalent mixture in which the 'control' RNA is the reference for expressing the gene transcript levels in the target sample (Figure 1).

mRNA is reverse transcribed into cDNA, which is then used to generate cRNA incorporating fluorescently labeled ribonucleotides. Generally, one sample is labeled with the Cy3 (green) fluorochrome, the other with Cy5 (red). The labeled target and reference cRNAs are equivalently mixed and hybridized together on the glass slide microarrays. The labeled cRNA molecules hybridize to the corresponding cDNA or long oligonucleotides, of which the exact position on the array is known. After scanning the slide with a laser-based scanner, computer calculations provide the ratios between green and red fluorescence for each spot, corresponding to the relative abundance of RNA from a particular gene in the target sample *vs* the reference sample. Computer-based spot recognition, normalization and ratio determination are used.

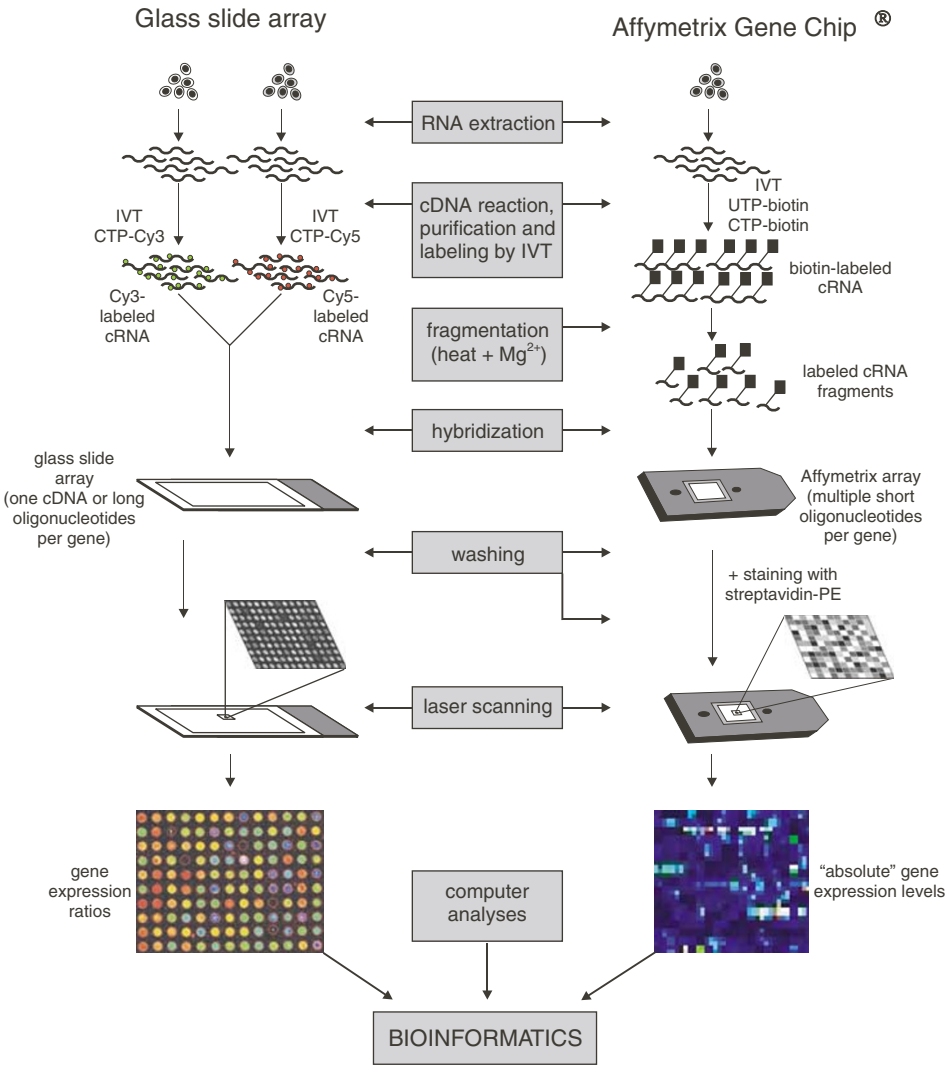

Figure 1 Comparison of glass slide and Affymetrix microarray procedures. For glass slide experiments, two cell populations, for instance diseased and normal, are isolated, RNA is extracted, and cDNA is made, which is used for *in vitro* transcription (IVT) with Cy3 (green) or Cy5 (red) labeled nucleotides. The two labeled cRNA samples are mixed and hybridized on a glass slide array, which is scanned with a laser, followed by computer analysis of the intensity image. With Affymetrix arrays, one population is used as starting material. Total RNA is extracted and cDNA is prepared. The cDNA is used in an IVT reaction to generate biotinylated cRNA. After fragmentation, this cRNA is hybridized to microarrays, washed and stained with PE-conjugated streptavidin, and subsequently scanned on a laser scanner.

Two different subtypes of glass slide microarrays can be discerned, based on the use of cDNAs or oligonucleotides. The cDNA microarrays have been around for some time and are often produced by spotting PCR products on glass, later transformed into single-strand (ss) DNA products by treatment with alkali or light. Given the problems with reproducible spotting of products of different lengths and generating ssDNA products, glass slides with oligonucleotides have been produced. These oligonucleotides are typically 70 mers or 80 mers. This is much longer than the 25 mers used by Affymetrix, but generally only one oligonucleotide is used per gene without any hybridization control.

Glass slide microarray technology is readily amendable for relatively large numbers of samples, not as expensive as Affymetrix technology, and can be set up by investigators themselves using an array spotter. However, the technical difficulties in the reproducible production of one's own glass slide microarrays should not be underestimated: conditions such as moisture, temperature, and light intensities in the room of the spotter need to be carefully controlled, large numbers of oligonucleotides of similar quality need to be present, and in the case of cDNA glass slides roughly similar hybridization conditions for each cDNA need to be found.

Affymetrix microarrays

Affymetrix microarrays (the so-called GeneChips[7]) generate a gene expression profile of one sample and therefore use only one color (phycorerythrin, PE, red). The design of the GeneChips is such that expression of a gene is interrogated by several (11–20) 25-mer oligonucleotides that span a part of the gene (Figure 2). In addition to these perfect-match oligonucleotides, each 25 mer comes with a negative control oligonucleotide that contains a mismatch at position 13. The integration of the expression levels for each of the 11–20 perfect-match–mismatch oligonucleotide sets generates a value for the expression of a particular gene.

Comparable to the glass slide microarrays, mRNA is reverse transcribed into cDNA, which is subsequently used as a template in an *in vitro* transcription reaction that incorporates biotinylated ribonucleotides into the cRNA (Figure 1). The cRNA is hybridized to the 25-mer oligonucleotides on the GeneChip and is subsequently stained with streptavidin–PE, followed by an optional extra stain with a PE-conjugated anti-streptavidin antibody for signal enhancement. After the washing and staining procedures, the GeneChip is scanned with an Agilent Gene Array laser. The scanning results are analyzed with Affymetrix proprietary software and an expression profile is generated. This technology offers several advantages over glass slide technology (see Table 1). Firstly, it is well suited for comparisons of multiple samples because no ratios are used, making it a natural platform for large series of clinical samples without the need of pairwise analyses. Secondly, the high reproducibility between arrays makes comparisons between separately analyzed samples more reliable. Thirdly, for most applications glass slides require more RNA than Affymetrix arrays, which can be problematic, particularly for diagnostic research with patient samples. Major disadvantages of the Affymetrix technology are its high price and the impossibility of making one's own arrays, although Affymetrix has overcome this problem to some extent by offering (expensive) custom arrays.

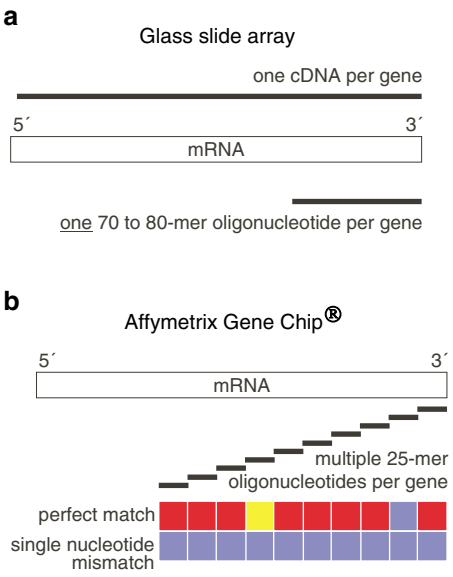

Figure 2 Design of Affymetrix microarrays (GeneChip) and long oligonucleotide arrays. While Affymetrix employs 11–20 different and sometimes overlapping 25-mer oligonucleotides per gene, a glass array uses a single cDNA or long oligonucleotide per gene. In addition to perfect-match oligonucleotides, Affymetrix GeneChips also contain mismatch oligonucleotides that serve as negative controls. The mismatch oligonucleotides carry a mutation at position 13 of the 25 mers. Intensity images of perfect match (upper row of squares) and mismatch (lower row) are indicated. In cases where both perfect match and mismatch yield a strong signal, the contribution of that probe to the overall expression is ignored, because it is not specific.

Bioinformatics for analysis of microarray results

Whatever the platform used, after generation of the expression profile of the samples, a major challenge is put forward by the analysis of the data. This forms the realm of bioinformatics and cannot be taken lightly. As a first step, data are visualized with the aid of clustering techniques, such as hierarchical clustering and self-organizing maps. Additional steps that are typically performed include filtering (removal of genes that are not expressed), normalization (eg to expression levels of house keeping genes or to the overall expression level), comparison of samples, for example, to identify new marker genes or new

Table 1 Comparison of various microarray platforms

Platform	*cDNA glass slide*	*Long oligonucleotide glass slide*	*Affymetrix GeneChip*
Design	cDNAs or PCR products spotted on glass slide	70- or 80-mer ss DNA oligonucleotides spotted on glass slide	Series (11–20) of 25-mer oligonucleotides with mismatch controls, synthesized on silicone chip
Sensitivity	One in 50 000–100 000	One in 300 000	One in 300 000
Minimal amount of RNA needed	10–50 μg	5–10 μg	3–5 μg
Advantages	Easy to produce Flexible to include new genes	Relatively easy to produce Flexible, reproducible	Reproducible Reliable data because of multiple probe design Standard platform allows comparison with data from others
Disadvantages	Reproducibility sometimes problematic Poor signals due to inefficient generation of ssDNA products Comparisons with other researchers problematic (no standard format)	Quality of oligonucleotides can be problematic Comparisons with other researchers problematic (no standard format)	Expensive Impossible to make one's own arrays
Typical applications	Comparisons of two samples	Comparisons of two samples	Comparisons of multiple samples requiring individual data sets or requiring multiple analyses on each data set

Table 2 Characteristics of seven relapsed precursor-B-ALL patients

Patient	*Code*	*Age at diagnosis*	*Sex*	*Tumor load (before purification) %*	*Purification*	*Chromosome aberration*	*Diagnosis-relapse interval*
1	ALL-12-D	0y2mo	M	>95	—	t(4;11)	6mo
	ALL-13-R			>95	—		
2	ALL-16-D	12y4mo	M	87	CD34	t(9;22)	6mo
	ALL-17-R			85	CD34		
3	ALL-22-D	2y8mo	M	85	CD19	t(12;21)	6mo
	ALL-14-R			>95	—		
4	ALL-8-D	3y11mo	M	>95	—	t(12;21)	11mo
	ALL-9-R			>95	—		
5	ALL-20-D	5y5mo	M	50	CD19	t(12,21)	3y7mo
	ALL-21-R			>95	—		
6	ALL-18-D	8y7mo	M	60	CD19	t(12;21)	4y11mo
	ALL-19-R			30	CD19		
7	ALL-15-D	9y1mo	M	75	CD19	?	2y5mo
	ALL-23-R			60	CD19		

y, years; mo, months.

criteria for diagnosis and classification. This is followed by a data mining step, in which the biomedical literature and databases are searched for information about the genes found and about gene–gene relationships. A final step involves the integration of these data with other information sources, for example, clinical data (disease progression and final outcome), or data about biochemical pathways or signal transduction routes.

Strategies: sample selection and handling

Given the enormous amount of work and costs involved in microarray experiments, it is advisable to plan the study thoroughly and to use high-quality samples, which often necessitates a sample purification step.

To illustrate these considerations about microarray experiments, we here report on a pilot study on patients with a relapse of precursor-B-cell ALL. We employed the Affymetrix technology for the comparison of paired diagnosis and relapse samples of precursor-B-ALL patients. Although paired analyses (such as diagnosis–relapse comparisons) can in principle be performed with glass slide microarrays, the lack of individual expression profiles precludes further comparisons of all individual samples. As shown here, such further analyses can, for instance, be carried out to compare early *vs* late relapsing patients with the same chromosomal translocation. It is because of this possibility to compare individual expression profiles that we have chosen Affymetrix as a platform for this study. Moreover, we illustrate the need for purification of the leukemia cells in order to minimize false-positive and -negative rates. The identified genes may represent useful new markers for classification or disease progression in precursor-B ALL.

Material and methods

Patient selection

To perform our comparative gene expression profiling study on paired diagnosis–relapse samples from precursor-B-ALL patients, we started with selection of cell samples from ~100 relapsed ALL cases, including ~25 t(12;21) positive patients as assessed by the presence of *TEL-AML1* transcripts. These *TEL-AML1*-positive cases formed our original target group with specific focus on early *vs* late relapses.

Our earlier microarray studies showed that currently available RNA amplifications steps result in unwanted skewing of particular gene transcripts (Staal *et al*, unpublished results), implying that RNA amplification should be avoided in studies on gene expression profiling, unless better techniques become available. Furthermore, we observed in our first ALL gene expression analyses that several seemingly differentially expressed genes were actually derived from contaminating non-leukemic cells, particularly myeloid cells and T-lymphocytes (see Results). Consequently, for optimal results we decided to purify the ALL cell samples if the tumor load was <95% (see below).

Despite the availability of a relatively large series of paired diagnosis–relapse samples, we were confronted with the fact that only in four *TEL-AML1*-positive cases both the diagnosis and the relapse cell samples yielded sufficient RNA after thawing, purification of the ALL cells, and RNA extraction. Therefore, we broadened our criteria to include all cases phenotypically identified as common ALL, thereby having to include other translocations as well. In seven cases, sufficient RNA was obtained from both diagnosis and relapse samples. Characteristics of the seven selected patients are summarized in Table 2, including data on translocation, tumor load, and remission duration.

Purification of tumor samples

Based on the immunophenotype of the ALL samples, the most appropriate marker for purification of tumor cells was chosen. This was either CD19 or CD34. Tumor cells were purified with magnetic beads coated with CD19 or CD34 antibodies using standard procedures and MACS (Miltenyi, Bergisch Gladbach, Germany).

RNA extraction and DNA microarray procedures

RNA was isolated using RNeasy columns as described by the manufacturer (Qiagen, Hilden, Germany). The integrity of the RNA was tested on 1% formaldehyde containing agarose gels. A total of 3–5 μg of RNA was used to generate ds cDNA using superscript reverse transcriptase and a T7-oligodT primer. The

resulting cDNA was used in an *in vitro* cRNA reaction using T7 RNA polymerase and biotinylated ribonucleotides employing an ENZO kit (ENZO, Farmingdale, NY, USA). The biotinylated cRNA was cleaned-up using RNeasy spin columns (Qiagen) and quantified by spectrophotometric methods. An adjusted cRNA yield was calculated to reflect carryover of unlabeled total RNA. Fragmentation of 20 μg cRNA was performed at 95°C for 35 min. cRNA (5 μg) was hybridized to a Test3 microarray (Affymetrix) to check the quality of the procedure. Fragmented cRNA (10 μg) was subsequently hybridized for 16 h to U95A microarrays (Affymetrix) at 45°C. After washing and staining with PE-conjugated streptavidin, the arrays were scanned in an HP/Affymetrix scanner at 570 nm using a krypton/argon laser.

Bioinformatics

Data reported here are extracted from a total of 15 different U95Av2 or U95A GeneChips used under the various conditions described. The scanned images were analyzed using Affymetrix Microarray suite 4.2 software, using the diagnosis sample as baseline. Further analysis was performed using micro-DB and desktop Mining Tool 3.0 software. Statistical significance was tested using the freeware program SAM (significance analysis of microarrays) version 1.12. Normalization and correction of variance were carried out as described in detail in the Results section.

Ratios between the 5′ oligonucleotides and the 3′ oligonucleotides of GAPDH transcripts were <1.5 (usually 0.9–1.1.), indicating that the amount of labeling was equally distributed over the RNA molecules. This implies that no major degradation of RNA occurred. In comparison experiments care was taken that the scaling factor, noise, and presence calls were all comparable.

Results and discussion

Purification of cell samples

Strikingly, in one of our first experiments comparing diagnosis and relapse samples, a number of genes were identified as differentially expressed genes that normally would not be expressed in B cells, such as β and δ globin, *TCRB*, and neutrophil peptide gene 3 (data not shown). This indicated that transcripts from non-leukemic cells had 'contaminated' the RNA sample. This would obviously lead to false identification of genes to be used as prognostic or diagnostic markers. We therefore decided to investigate the problem of tumor load and contaminating cells that contribute to the expression profile.

We took a bone marrow sample with a relatively high tumor load of almost 80% at diagnosis and purified the tumor cells using MACS and CD19 beads. This increased the tumor cell contribution in the total cell sample to 95% or more (Figure 3). We also purified the relapse sample (Figure 3). RNA was extracted from three samples (unpurified diagnosis sample, purified diagnosis sample, and purified relapse sample), labeled and expression profiles were determined for the 12,600 genes present on U95A GeneChips. As shown in Figure 4, comparison of the expression profiles of the purified and non-purified diagnosis samples showed a wide distribution of several hundreds of genes that differed more than two-fold, potentially leading to the false identification of genes that seemed to differ between diagnosis and relapse of the same tumor (in the same patient). Indeed, comparisons of the expression profiles of the purified and non-purified diagnosis sample with the purified relapse sample demonstrated that many more genes would have erroneously been identified as genes up- or downregulated after relapse (Figures 4b and 4c). Using the non-purified diagnosis sample, 388 genes would have been found upregulated at relapse and 177 downregulated (two fold or more), compared to only 104 and 88 with a purified sample, respectively. Therefore, 211 upregulated and 89 downregulated genes would have been falsely identified. For some of these genes (such as β and δ globin or *TCRB*) this may be obvious, but many other genes are not known to be associated with non-B-cell lineages and are therefore hard to detect as 'contamination'.

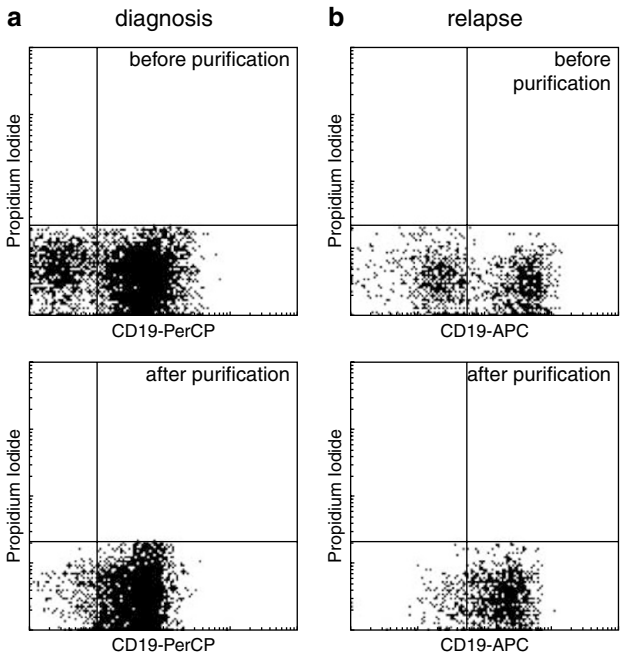

Figure 3 Purification of precursor-B-ALL cells. FACS plots for CD19 *vs* propidium iodide at diagnosis (a) and relapse (b), each before and after purification with CD19 beads and MACS.

We conclude that purification of the tumor sample is essential to avoid contamination of normal cells and thereby to obtain meaningful gene expression profiling. Therefore, we purified every ALL sample in our study when the tumor load was less than 95%. The reason for choosing a cutoff level of 95% instead of 90 or 85% was based on the finding that percentages of 10–15% of non-leukemic cells already give misleading results and because $>95\%$ purity can routinely be obtained by purification via magnetic beads, although cell loss after thawing and magnetic bead purification might be substantial. Based on the immunophenotype of the involved precursor-B ALL, we have used either CD19 or CD34 magnetic beads for the purification of the leukemic cells. Owing to our strict criteria for purity of the ALL samples, we could analyze only a limited number of ALL cases.

Normalization

It is the purpose of the data analysis process to detect genes that are consistently differentially expressed in the two sets of samples, that is, genes that are up- or downregulated in the diagnosis (*D*) group with respect to the relapse (*R*) group and

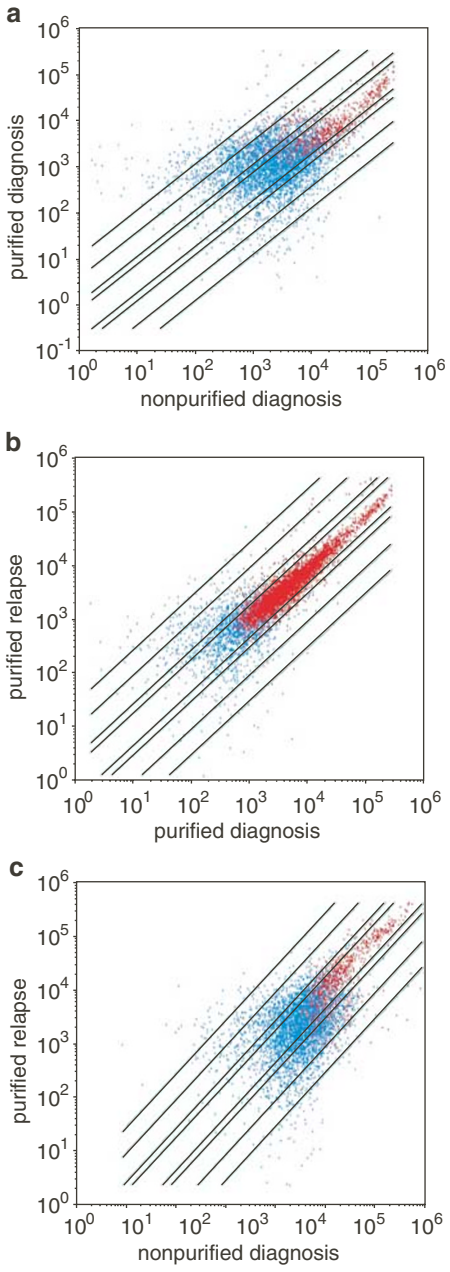

Figure 4 The need of purification of leukemic cells as demonstrated by gene expression profiling. (a) Comparison of gene expression profile of CD19-purified and nonpurified precursor-B-ALL samples of the same patient. (b) Comparison of diagnosis and relapse sample of this patient, with both diagnosis and relapse purified. (c) identical to panel b, but without purification of the diagnosis sample.

vice versa. In determining whether a gene is consistently differentially expressed, the difference in expression level in the pair of arrays originating from the same patient is examined, for all patients. Suitable statistical tests (eg Wilcoxon's matched-pairs signed-rank test) exist to perform such a 'matched-pairs' analysis. However, before such tests can be applied, proper normalization between arrays must be performed. If this is not done, variations between arrays – not related to the biological phenomena – might be detected by the tests as significant variations in expression. For example, if the mean expression

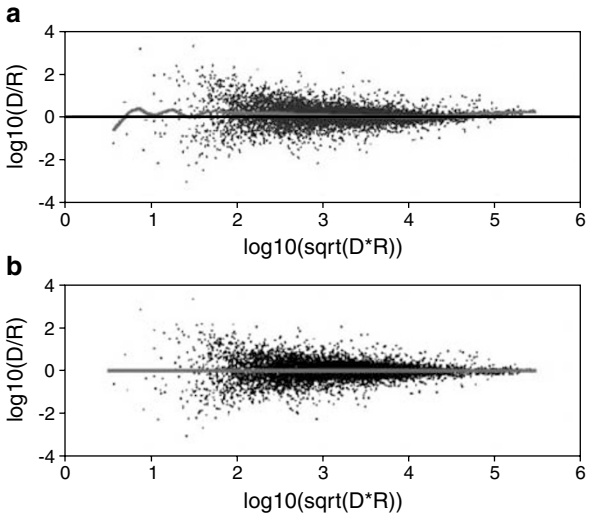

Figure 5 Normalization strategy. The intensity–ratio plot for the *D*/*R* log ratio for a particular patient. In panel a, the uncorrected log ratio is displayed with the solid line denoting the estimated intensity-dependent average log ratio. Panel b shows the ratios after correction of the mean ratio to a log ratio of zero.

value (per array) in all arrays in set *R* is higher than the mean expression values in set D, then all truly unchanged genes will show a significant downregulation from *D* to *R*. Since we were specifically interested in expression changes within each *(D/R) pair,* we employed a normalization strategy where each *pair* was normalized to correct for any array effects.

To correct for effects between arrays present in the *D*/*R* ratio, we employed the same approach as employed to normalize the red/green ratio in cDNA arrays.[8] This normalization process is based on two assumptions concerning the *D*/*R* ratio of all genes in a matched pair: (1) the majority of genes have a log ratio of zero, that is, unchanged expression, and (2) the number of up- and downregulated genes is roughly equal. Consequently, the normalization process consists of two steps: (1) correcting for the intensity-dependent offset in the ratios, that is, forcing the average log ratio as a function of intensity to a value of zero and (2) correcting for the intensity-dependent variation in the variance of the log ratio.

Step 1: *Ratio offset correction*. Figure 5 depicts the intensity–ratio plot for all genes of a particular matched pair. The horizontal axis represents a variable proportional to the intensity in both arrays, while the vertical axis represents the log ratio, *D*/*R*. As expected, the log ratio is around (but not exactly) zero for the whole intensity range. The wavy solid line in the top panel represents the kernel-smoothed estimate of the log ratio as a function of intensity. Then the log ratio for every gene is corrected by adjusting the log ratio such that the average intensity-dependent log ratio is zero everywhere. In the lower panel, the resulting corrected cloud of ratio data points is shown. The average log ratio is now a straight line equaling zero for the whole intensity range.

Step 2: *Ratio variance correction.* The clouds in Figure 6 exhibit an 'arrowhead' shape: large log-ratio variations (around zero) for low intensities and small variations for large intensities. This effect is corrected for in the second step of the normalization process, such that a fair comparison of genes expressed at different intensities is possible. First, the standard deviation of the log ratio (around the log ratio of zero) is estimated as follows. The intensity range is divided in equally sized bins and the

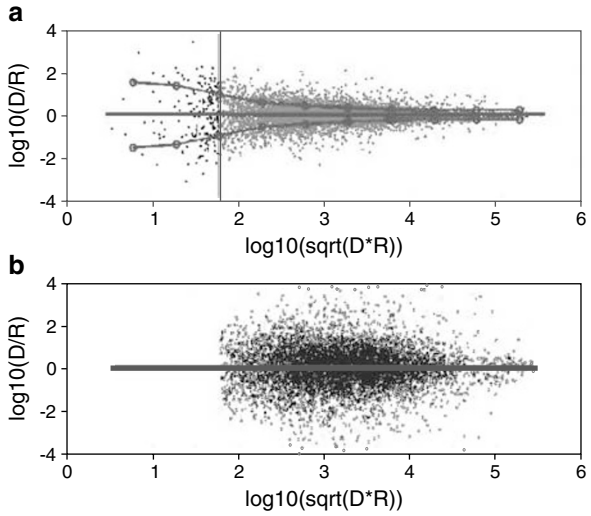

Figure 6 Graphical representation of Step 2 of the normalization process. The top panel shows the data cloud with the estimated standard deviation in the log ratio as a function of the intensity indicated by the solid line connecting the open dots. The vertical line indicates the minimal intensity employed in the final analysis. Panel b shows the data cloud after correction of the standard deviation and removal of the low-intensity data points.

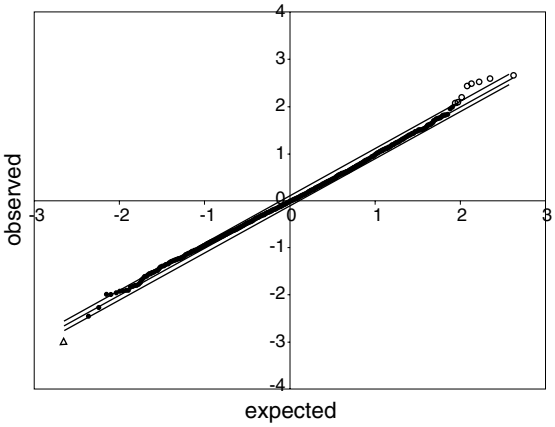

Figure 7 SAM plot for diagnosis-relapse comparison on seven common ALL patients. Data for the 14 chips were normalized and subjected to SAM analysis. Differentially expressed genes are indicated by open symbols. Eight positively (but only five at a *q*-value of 16.6, three at a *q*-value of 56; open circles) and one negatively regulated gene (open triangle, *q*-value = 16.6) were found.

Table 3 Genes that significantly differ between relapse and diagnosis precursor-B ALL (SAM analysis)

Name (annotation)	*Q-value**	*Number of pairs up/down regulated*	*Up- or down regulated in relapse*
SH3 BP5	16	7/7	+
TOSO	16	7/7	+
Smad3	16	7/7	+
CD79a (Mb1)	16	7/7	+
RAS GRP2	16	7/7	+
WAS F1	16	7/7	−

*The *Q*-value represents a measure of significance in a SAM test.

standard deviation of the log ratio is estimated in each bin. The open circles in Figure 6 represent these estimates. For bins with too few genes particularly at the low end of the intensity scale (near the detection limit), this estimate becomes unreliable. Therefore, these genes were discarded. The boundary below which genes are discarded is indicated with a vertical line in Figure 6a. This cutoff point was determined by a significant (100-fold) change in the number of genes within a bin. After computing the estimates of the standard deviations, the log ratio of each gene is scaled by the standard deviation associated with the intensity for that gene. For example, genes with a high intensity have a small standard deviation and are, therefore, scaled such that the standard deviation increases with respect to the low-intensity genes. The result is depicted in Figure 6, panel b.

Identification of disease progression markers

In order to avoid too stringent testing, such as step-down correction,[9,10] we employed the SAM package.[11] SAM is a statistical method specifically designed for identification of differences in large-scale microarray-based gene expression profiles. SAM computes a statistic for each gene (in this case the ratio of the mean *D*/*R* log ratio) and standard deviation across all pairs. Then the 'false discovery rate' (FDR) is estimated by analyzing permutations of the measurements. This rate expresses the percentage of genes identified as significant, purely by chance for a given value of the threshold parameter, Δ.

We determined the expression profiles of seven relapsed common ALL patients at diagnosis and at relapse. Between samples of one individual, the number of genes that differed between diagnosis and relapse might be as high as 80 upregulated and 80 downregulated genes, based on the popular two-fold change in expression per gene. While this is a common cut-off value, there is no statistical or biological basis for this value. We first identified which genes differed between diagnosis and relapse per patient and by how much. Further application of SAM analysis (Δ set such that the FDR is 16%) yielded five significantly upregulated and a single significantly downregulated gene (Figure 7 and Table 3). These differentially regulated genes were significantly up- or downregulated in all seven patients.

Most of the identified genes that were upregulated encode signaling molecules: the Btk-associated SH3-domain containing protein SH3 BP5, the Ras guanyl releasing protein RASGRP2, and the TGFβ responsive transcription factor Smad3. These markers are associated with increased proliferation and responsiveness to proliferative signals in the relapse samples. This is further supported by the upregulation of the antiapoptotic gene *TOSO*. Toso functions as an inhibitor of Fas-mediated apoptosis in hematopoietic cells via inhibition of caspase 8 processing.[12] Obviously, such a gene will give a growth advantage to the leukemic cells. This protein might therefore be an attractive drug target.

Another marker found to be upregulated, the immunoglobulin (Ig)-associated gene CD79A (mb-1), is present at moderate levels in immature B cells (seen as the normal counterpart of precursor-B-ALL cells) and at increased levels in more mature B cells. This would either indicate differentiation *in vivo* to more mature stages or, more likely, that in some patients subclones

with a 'more mature' precursor-B-ALL phenotype more resembling pre-B ALL came up at relapse that were not readily detected at diagnosis. Such emersion of subclones has been shown before using RQ-PCR of the rearranged Ig genes.[13,14]

The single statistically significant downregulated gene *WASF1* concerns a member of the Wiskott Aldrich syndrome protein (WASP) gene family that is involved in actin filament reorganization and changes in cell shape. WASP expression is regulated by the small GTPases Rac and cdc42, which are related to Ras. Given that a Ras-associated gene also changes in expression, it is possible that regulation of small GTPases is altered in relapse samples.

Interestingly, no multidrug-resistance markers (eg MDR-1) or genes involved in xenobiotic metabolism (eg P450, GSH transferases) were found. It is striking that no such markers were found, although four out of seven patients relapsed *during* treatment. Given the small number of patients in our study, it is not possible to conclude whether the absence of drug-resistance markers is a consistent finding in all early relapsed patients.

It is of interest that among the seven precursor-B-ALL patients studied, the four patients with t(12;21) showed different remission durations. Two patients relapsed early (<1 year, 5th percentile of relapses in precursor-B ALL), whereas two others relapsed late (85th percentile). In addition, the two patients with translocations associated with poor prognosis (t(9,22) and t(4,11)) could be analyzed as a third separate group. Comparison of the expression profiles at diagnosis between these three groups of patients could in principle identify gene clusters that predict differences in clinical outcome. For this kind of analysis, it is imperative to investigate individual samples. This possibility is an attractive feature of Affymetrix-based microarrays. In the current study, this would mean a comparison between the three sets of two patients, which we have done here for the comparison of early *vs* late relapsing t(12;21) positive patients. However, this comparison did not yield statistically significant results (FDR = 64%; data not shown). It therefore remains to be determined in a larger cohort of patients whether a gene cluster can be identified that constitutes a signature profile for early relapse and could serve as a clinical predictor. The here described bioinformatical considerations should also be applied to such a larger cohort. These types of analyses are ideally done on platforms that yield absolute expression profiles rather than ratios, since besides the initial pairwise comparisons other types of analyses are needed, which require reinvestigation of the data per individual patient.

Conclusion

Although our microarray study is small, it is the first one on comparison of diagnosis and relapse samples, for which sufficient cell material was available to extract RNA, even after purification of the leukemic cells. Comparative pairwise diagnosis–relapse studies have the advantage over classification studies that potential new markers and drug targets might be found in relatively small patient series. In addition to these comparisons, identification of subgroups and markers that predict outcome at diagnosis is also a highly valuable approach. The latter type of study needs many diagnosis samples of which clinical outcome is known. A diagnosis–relapse study as we report here will identify a different kind of marker that may be informative for disease progression.

In conclusion, our small study on the comparison of paired diagnosis and relapse in precursor-B ALL by microarray analysis identified genes that are associated with cell proliferation and cell survival, but not with drug resistance, although four patients relapsed during treatment. These genes or their protein products are currently not in use as diagnostic or disease progression markers and represent potential new markers for such applications. Another conclusion from this study underscores the need for purification of the patient sample in order to prevent false-positive gene expression. In large published microarray studies this has not been performed thus far. It may be that in very large patient series (>100 patients), purification is not necessary because erroneously identified genes may be eliminated by the statistics of large numbers. However, it is striking that in a recently published study the T-cell-specific gene CD3ζ was identified as a gene predicting outcome in the hyperdiploid precursor-B ALL.[15] It would be interesting to see whether precursor-B ALL cells indeed express CD3 or whether contaminating T-lymphocytes were contributing to the expression profile. Lastly, some of the genes identified here may also provide potential drug targets to improve survival rates for acute leukemias.

Acknowledgements

We thank members of the Molecular Immunology unit from the Department of Immunology for stimulating discussions. This work was supported in part by the Dutch Royal Society for Arts and Sciences (KNAW; FJTS), the Anna and Maurits de Cock Foundation (FJTS), the Bekales foundation (FJTS) and The Dutch Heart Foundation (CvdB).

References

1 Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC. Distinct types of diffuse large B-cell lymphomaQ2 identified by gene expression profiling. *Nature* 2000; **403**: 503–511.
2 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**: 41–47.
3 Davis RE, Staudt LM. Molecular diagnosis of lymphoid malignancies by gene expression profiling. *Curr Opin Hematol* 2002; **9**: 333–338.
4 Ferrando AA, Neuberg DS, Staunton J, Loh ML, Huard C, Raimondi SC. Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell* 2002; **1**: 75–87.
5 Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; **286**: 531–537.
6 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**: 1639–1647.
7 Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. *Nat Genet* 1999; **21**: 20–24.
8 Yang YH, Dudoit S, Luu P, Speed TP. Q3*Normalization for cDNA microarray data*. San Jose: SPIE BiOS, 2001.
9 Westfall PH, Krishen A, Young SS. Using prior information to allocate significance levels for multiple endpoints. *Stat Med* 1998; **17**: 2107–2119.
10 Westfall PH, Zaykin DV, Young SS. Multiple tests for genetic effects in association studies. *Methods Mol Biol* 2002; **184**: 143–168.
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 Hitoshi Y, Lorens J, Kitada SI, Fisher J, LaBarge M, Ring HZ. Toso, a cell surface, specific regulator of Fas-induced apoptosis in T cells. *Immunity* 1998; **8**: 461–471.

13 Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 2002; **1**: 133–143.
14 De Haas V, Verhagen OJ, von dem Borne AE, Kroes W, van den Berg H, van der Schoot CE. Quantification of minimal residual disease in children with oligoclonal B-precursor acute lymphoblastic leukemia indicates that the clones that grow out during relapse already have the slowest rate of reduction during induction therapy. *Leukemia* 2001; **15**: 134–140.
15 Szczepanski T, Willemse MJ, Brinkhof B, van Wering ER, van der Burg M, van Dongen JJ. Comparative analysis of Ig and TCR gene rearrangements at diagnosis and at relapse of childhood precursor-B-ALL provides improved strategies for selection of stable PCR targets for monitoring of minimal residual disease. *Blood* 2002; **7**: 2315–2323.