Gene expression–based classification and regulatory networks of pediatric acute lymphoblastic leukemia

*Zhigang Li,[1] *Wei Zhang,[2,3] Minyuan Wu,[1] Shanshan Zhu,[2,3] Chao Gao,[1] Lin Sun,[1] Ruidong Zhang,[1] Nan Qiao,[2,3] Huiling Xue,[2] Yamei Hu,[1] Shilai Bao,[4] Huyong Zheng,[1] and Jing-Dong J. Han[2]

[1]Beijing Children's Hospital of Capital Medical University, Beijing; [2]Chinese Academy of Sciences Key Laboratory of Molecular and Developmental Biology, Center for Molecular Systems Biology, Beijing; [3]Graduate School, Chinese Academy of Sciences, Beijing; and [4]Chinese Academy of Sciences Key Laboratory of Molecular and Developmental Biology, Center for Molecular Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

Pediatric acute lymphoblastic leukemia (ALL) contains cytogenetically distinct subtypes that respond differently to cytotoxic drugs. Subtype classification can be also achieved through gene expression profiling. However, how to apply such classifiers to a single patient and correctly diagnose the disease subtype in an independent patient group has not been addressed. Furthermore, the underlying regulatory mechanisms responsible for the subtype-specific gene expression patterns are still largely unknown. Here, by combining 3 published microarray datasets on 535 mostly white children's samples and generating a new dataset on 100 Chinese children's ALL samples, we were able to (1) identify a 62-gene classifier with 97.6% accuracy from the white children's samples and validated it on the completely independent set of 100 Chinese samples, and (2) uncover potential regulatory networks of ALL subtypes. The classifier we identified was, thus far, the only one that could be applied directly to a single sample and that sustained validation in a large independent patient group. Our results also suggest that the etiology of ALL is largely the same among different ethnic groups, and that the transcription factor hubs in the predicted regulatory network might play important roles in regulating gene expression and development of ALL. (Blood. 2009;114: 4486-4493)

Introduction

Acute lymphoblastic leukemia (ALL; aka, acute lymphocytic leukemia or acute lymphoid leukemia) is the most common malignancy diagnosed in children, representing nearly one-third of all pediatric cancers, with a peak incidence in 2- to 5-year-old children.[1]

ALL is a heterogeneous disease with more than 12 subtypes that respond differently to chemotherapy.[2-9] If a patient is correctly classified into a specific risk group, and treated with corresponding aggressiveness so that the patient is neither overtreated nor undertreated, the cure rate can exceed 80%.[6,9,10] Therefore subtype classification is very important in ALL diagnosis. The 6 common subtypes of ALL are T-ALL, t(1;19) (*E2A-PBX1*), t(12;21) (*TEL-AML1*), t(9;22) (*BCR-ABL*), t(4;11)*MLL*-rearrangement, and hyperdiploid with more than 50 chromosomes (hyperdiploid>50; Table 1).

Morphology, immunology, cytogenetics, and molecular biology classification is widely used clinically for pediatric ALL. However, as it is an expensive and time-consuming process; it is available only in developed countries and a few major medical centers in some developing countries. DNA microarrays have spurred the search for gene expression–based markers for computational ALL classification. By comparing genome-wide gene expression among the subtypes of ALL, approximately 80 to 300 genes have been identified as marker genes necessary to discriminate the 6 major subtypes.[11,12] However, a classification model that can be applied to a single independent patient sample and can consistently retain high accuracy is still lacking. Furthermore, the molecular mechanisms giving rise to the subtype-specific gene expression patterns are poorly understood. This led us to investigate whether by combining more gene expression profiles from different studies we can (1) find a minimal general set of marker genes for clinical ALL subtype classification, and (2) uncover the possible regulatory networks of ALL subtypes.

Methods

Datasets

White children's ALL datasets were obtained from Yeoh et al,[11] Ross et al,[12] and Hoffman et al.[13] Gene annotations were downloaded from Gene Ontology[14,15] (GO; http://www.geneontology.org) and Kyoto Encyclopedia of Genes and Genomes[16,17] (KEGG; http://www.genome.jp/kegg). Transcription factor (TF) binding motifs were from TRANSFAC[18,19] and JASPAR[20,21] databases. Predicted functional interactions were derived from IntNetDB.[22,23] Drug targets were obtained from Yildirim et al.[24] All studies were approved by the institutional ethical board of the Chinese Academy of Sciences.

CEL file preprocessing

The CEL file for each white or Chinese sample was preprocessed individually using 2 steps of robust multiarray analysis (RMA)[25] from the

Submitted April 24, 2009; accepted July 30, 2009. Prepublished online as *Blood* First Edition paper, September 15, 2009; DOI 10.1182/blood-2009-04-218123.

The online version of this article contains a data supplement.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

*Z.L. and W.Z. contributed equally to this work.

Table 1. Six major subtypes of ALL

Subtype	Occurrence, %	Clinical character
t(9;22)(*BCR-ABL*)	2-3	High risk
t(1;19)(*E2A-PBX1*)	5	Low risk
t(12;21)(*TEL-AML1*)	16-22	Normal ALL low risk
t(4;11)(*MLL*)	5-8	Infant ALL high risk
Hyperdiploid>50	25-35	Normal ALL low risk
T-ALL	10-13	T-ALL moderate risk

ALL indicates acute lymphoblastic leukemia.

Bioconductor package[26]: (1) RMA convolution background adjustment and (2) summarization based on "multiarray model" using the "median polish" algorithm.

Data normalization between different microarray platforms

To reduce the systematic differences between Affymetrix HG-U95A and HG-U133A, we first mapped all the probe sets on the Affymetrix HG-U95Av2 chip onto the HG-U133A chip probe set IDs using the HG-U95 to HG-U133 Best Match table (http://www.affymetrix.com/support/technical/byproduct.affx?product = hgu133). Only probe sets that mapped to unique genes were used for marker selections. We then used Ross et al's dataset as a standard to transform the mean and variance for each gene in Yeoh et al's dataset by the formula $x_i' = ([x_i - \mu]/\sigma \times * \sigma_0) + \mu_0$, where μ and σ are the mean and the standard deviation of the expression values of a gene in Yeoh et al's dataset, and μ_0 and σ_0 are the mean and the standard deviation of the gene's expression values in Ross et al's dataset. Ross et al's data, Hoffman et al's data, and our new Chinese array data, which are all based on HG-U133A, were not normalized. No other normalization was done in all the subsequent analysis steps. All microarray data have been deposited with Gene Expression Omnibus (GEO) under accession number GSE17703.[27]

Searching for the best minimal set of marker genes

We selected classification marker genes using the Support Vector Machine–Recursive Feature Elimination (SVM-RFE) tool.[28] Accuracy for each marker gene set was evaluated using 10-fold cross-validation for subtype classifier selection. We used 2 steps to select the minimal number of marker genes: (1) We ranked within each sample the 9116 single gene probe sets that are common to all datasets and used these rank values as initial input. We first eliminated 10% genes at each iteration while the number of remaining genes was more than 100, and then eliminated 1 gene at each iteration step while the number of remaining genes was less than 100. For each new iteration, the genes selected by the previous step were reranked within each sample as input. The criterion for a "candidate gene set" was the least number of genes that gave a 10-fold cross-validation accuracy greater than 93% for classification marker selection. (2) Then we combined the candidate genes of each fold (group) and further reduced the marker genes one by one based on their occurrence frequency as markers selected in each fold. The criterion for this "minimal marker set" was the least number of genes where the cross-validation accuracy is greater than 97% for classification marker selection.

The pseudocodes for selecting the minimal set of classification markers are provided in supplemental Materials (available on the *Blood* website; see the Supplemental Materials link at the top of the online article).

Classification using the 6 binary classifiers

When a sample was judged by the combination of the 6 SVM binary classifiers, if none of the 6 classifiers gave a positive result, we predicted the sample to be the "others" subtype; occasionally, if more than one classifier declared a positive result, we chose the classifier that placed the sample at the maximal distance to optimal separating hyperplane as the best classifier. The distance was calculated using the function $g(x) = (w \times x) + b$, where the vector w is the weights of the marker genes, the input vector x is the expression values of the marker genes, and b is a bias value.

Functional enrichment

Enriched KEGG pathways and GO terms were calculated as described by Xia et al.[29]

Chinese ALL samples

A total of 100 pediatric acute lymphoblastic leukemia bone marrow (BM) samples were analyzed, together with 5 non-ALL BM samples as negative control. The diagnosis of ALL was based on morphology, immunology, cytogenetic, and molecular classification. Cytogenetic ALL subtypes were identified experimentally by G-banding karyotype and multiplex nested reverse-transcription–polymerase chain reaction (PCR). Among the 100 ALL patients, 11 relapsed within 5 years. All the samples, including those relapsed afterward, were from patients treated on the Beijing Children's Hospital-2003 protocol and were extracted at their initial diagnosis. The 5 non-ALL BM samples were taken from the removed bones of patients who had plastic surgery for their bone deformity in Beijing Children's Hospital. And the informed consent was obtained from parents, guardians, or patients (as appropriate) in accordance with the Declaration of Helsinki. The detailed descriptions of these samples are provided in supplemental Table 1.

Gene expression profiling

Total RNA was extracted from cryopreserved mononuclear cell suspensions from BM samples using Trizol (Invitrogen) and purified with RNeasy Kit (QIAGEN). All samples were strictly subjected to quality control for RNA qualification. After extraction, the quality and quantity of RNAs were tested by electrophoresis and spectrophotometer, respectively. Only RNAs that had clear 28S and 18S bands on electrophoresis gels were used for microarray hybridization. Most samples had more than 20 μg total RNA and concentrations of 1.0 to 5.0 μg/μL (supplemental Table 2). cDNA and cRNA were synthesized with One-Cycle Target Labeling and Control Reagents (Affymetrix). The labeled RNA was then fragmented and hybridized to HG-U133A 2.0 oligonucleotide arrays (Affymetrix Incorporated) according to Affymetrix protocols. Using the Affymetrix Power Tools Package (http://www.affymetrix.com/partners_programs/programs/developer/tools/powertools.affx), the qualities of all microarray data were evaluated by the proportion of present calls and 3′/5′ intensity ratios of glyceraldehyde-3-phosphate dehydrogenase/ACTIN derived from the array intensity data (supplemental Table 2). All arrays with more than 20% present calls were included in the analysis regardless of 3′/5′ intensity ratios of glyceraldehyde-3-phosphate dehydrogenase/ACTIN. The CEL file of each sample was preprocessed individually using robust multiarray analysis (RMA) as described in "CEL file preprocessing." The CEL files for these microarray data are available at http://www.bch.com.cn/xy/BCH_ALL_microarray_data.rar.[30]

Pooling of 5 normal bone marrow samples

Total RNA from the 5 non-ALL BM samples was extracted and purified as described in "Gene expression profiling." Then the total RNA samples were pooled together using equal amount from each sample. A total of 10 μg was used for a single microarray hybridization.

Differential gene expression

Differentially expressed genes were identified using the RankProd program (Bioconductor).[31]

Text mining

We searched the PubMed abstracts for the co-occurrence of the genes with the term "leukemia" or a subtype name (eg, "BCR-ABL"). To test the significance of co-occurrence between a set of genes with the disease term, we randomly selected 100 sets of genes of the same number and from the same gene pool where the real gene set was obtained. After 100 such simulations, the empiric *P* value was taken as the number of simulations that had equal or more co-occurrences than the real gene set.

Quantitative PCR to validate gene expression measurements by microarrays

Total RNA were extracted using Trizol (Invitrogen) and reverse transcribed with RevertAid First Strand cDNA Synthesis Kit (Fermentas) according to the manufacturers' instructions. The gene expression levels were quantified with SYBR Green Real Time PCR Master Mix (TaKaRa) on iQ5 Real-Time PCR Detection System (Bio-Rad). ACTIN was used as internal control. The primers used are listed in supplemental Table 3. The expression level of each gene in the leukemic samples was quantified as a ratio relative to the average expression level of the gene in 5 normal samples. All RNAs for quantitative PCR (qPCR) were from the same patient samples as in the microarray assays but isolated independently from the cryopreserved mononuclear cells.

Identifying TF-binding motifs for each group of differentially expressed genes

We used the STORM software[32] (CREAD) to identify TF binding motifs based on the position weight matrix from the TRANSFAC[18,19] and JASPAR[20,21] databases within the 1-kb sequences upstream of transcription start site of the differentially expressed genes (obtained from University of California Santa Cruz hg18[33]), using the 1-kb sequences upstream of the transcription start sites of 1000 randomly selected genes as background. A P value less than .00001 was used as the criterion for the presence of a motif.

Results

Marker gene selection for 6 major ALL subtypes

We first investigated whether we can improve the subtype classification using more samples and a different marker selection method. We collected 3 ALL microarray datasets with a total of 535 samples (supplemental Table 4). The first dataset produced using Affymetrix HG-U95Av2 contains 335 samples.[11] The other 2, produced using Affymetrix HG-U133A microarray, contain 132 and 68 samples.[12,13] All the samples in these datasets have been experimentally classified into 6 known subtypes and 1 unknown (others) subtype. Our goal is to find the minimal number of marker genes to assign a sample to its determined subtype with the maximal accuracy.

Many methods can be or have been used to build a classifier, for example, Support Vector Machine (SVM),[34,35] Prediction by Collective Likelihoods,[36] decision tree, k-Nearest Neighbor (k = 1),[37] Naive Bayes,[38] and so on. However, candidate marker genes have to be selected by some arbitrary cutoff, such as a Student t test or χ^2 test P value, before their expression profiles can be used to train the classifiers. A recent improvement of SVM, named Support Vector Machine–Recursive Feature Elimination (SVM-RFE),[28] circumvents this problem by recursively selecting the most important features/genes for classification while running the SVM classifier, in addition to constructing classification models like other machine learning algorithms. Because of this major advantage of SVM-RFE over the other machine learning methods, we used it to build our classifiers.

For clinical diagnosis, each sample needs to be analyzed, classified, and diagnosed individually without many other samples being analyzed in parallel. To maximally simulate the clinical diagnosis setting, samples were analyzed individually instead of all together to minimize the statistical background reduction effect of large sample size.

We labeled the samples of a subtype "positive" compared with the rest of the "negative" samples that do not belong to the subtype.

Table 2. Cross-validation accuracies of our 62-gene classifier for each subtype on the 535 white children's samples

Subgroup	Accuracy, %	Sensitivity, %	Specificity, %
BCR-ABL	99.8	97.1	100
E2A-PBX1	100	100	100
Hyperdiploid>50	98.3	92.9	99.5
MLL	99.8	97.9	100
Others	97.6	97.3	97.6
T-ALL	100	100	100
TEL-AML1	99.6	99.0	99.8

Total accuracy = 97.6%. Sensitivity = true positive/(true positive + false negative). Specificity = true negative/(true negative + false positive).

We then used MSVM-RFE,[39] an extended version of SVM-RFE, for multiclass classifications, to find the genes that best separate each subtype from the rest of the samples and to construct classifiers. We first normalized Yeoh et al's data, which are on a different platform (Affymetrix HG-U95A) from that of all the other datasets (HG-U133A), based on Ross et al's data ("Data normalization between different microarray platforms" in "Methods"). Then, starting with a pool of 9116 probe sets, which could be mapped to unique genes and were common between HG-U95A and HG-U133A, we used the within-sample expression intensity ranks of the genes selected by the previous iteration as input values to construct the classifiers ("Searching for the best minimal set of gene markers" in "Methods"). This accommodates expression measurement variations between samples and the lack of statistical comparison during clinical diagnosis.

When constructing the classifier, we used 10-fold cross-validation to determine the accuracy of the classifier, which was defined as the fraction of correctly classified samples within all samples tested. The average accuracies over the 10 tests were used as criteria for classifier selection. We first selected the top 66 genes of each fold (group), which is the minimal number of genes required to achieve a 93% average accuracy among each of the 10 folds using SVM-RFE, and combined these genes as candidates ("Searching for the best minimal set of gene markers" in "Methods"). Then, based on the occurrence frequency, we further reduced the number of marker genes using the classification accuracy of 97% as a cutoff ("Searching for the best minimal set of gene markers" in "Methods"). The best classifier combination contains 62 genes (supplemental Table 5).

The classifier has an overall accuracy of 97.6% by 10-fold cross-validation among the 535 samples, with only 13 of the 535 samples misclassified (Table 2). This prediction accuracy is slightly higher than the classifiers found by Yeoh et al and Ross et al (96% on 335 samples and 97.2% on 132 samples, respectively), with similar number of marker genes to those used by Yeoh et al and much fewer than by Ross et al (120-300 genes).[11,12] It should be noted that the more samples and datasets included, the more difficult to achieve high cross-validation accuracy with few genes. So cross-validation accuracy tends to overestimate a classifier's performance on a small sample set. Hoffman et al have identified a predictor of 26 genes with a prediction accuracy of 98% on 104 samples without the others samples,[13] which do not have uniform expression profile and are the most difficult to predict (Table 2). Excluding the others samples, our classifier's prediction accuracy is 99%. The 62 marker genes are associated with the GO[14,15] (Gene Ontology) terms and the KEGG[16,17] (Kyoto Encyclopedia of Genes and Genomes) pathways related to leukocyte development and motility (Table 3, supplemental Table 5).

Table 3. Gene Ontology terms and Kyoto Encyclopedia of Genes and Genomes' pathways enriched among the 62 classification marker genes

Annotation type/ID	GO term/KEGG pathway	*P*	Fold enriched	Gene symbols
GO terms				
GO:0005178	Integrin binding	.002	20.04	*CTGF, ICAM3, ACTN1*
GO:0005902	Microvillus	.002	52.22	*CLIC5, PROM1*
GO:0003823	Antigen binding	.002	21.54	*IGJ, LILRA2, IGHD*
GO:0005834	Heterotrimeric G-protein complex	.004	23.93	*GNG11, GNAI1*
GO:0005520	Insulin-like growth factor binding	.005	24.98	*CTGF, IGF2R*
GO:0009611	Response to wounding	.005	24.98	*CTGF, MDK*
GO:0051015	Actin filament binding	.007	16.90	*MARCKS, ACTN1*
GO:0006396	RNA processing	.001	12.77	*IGF2BP3, RBMS1*
KEGG pathway				
04810	Regulation of actin cytoskeleton	.002	4.78	*WASF1, ITGA6, PIK3R3, ACTN1, ARHGEF4*
04662	B-cell receptor signaling pathway	.002	9.47	*JUN, PIK3R3, BLNK*
04510	Focal adhesion	.003	5.10	*ITGA6, COL6A3, JUN, PIK3R3, ACTN1*
04512	ECM-receptor interaction	.004	6.86	*ITGA6, COL6A3, FNDC3A*
04660	T-cell receptor signaling pathway	.004	6.41	*JUN, PIK3R3, ZAP70*
04670	Leukocyte transendothelial migration	.006	5.14	*PIK3R3, ACTN1, GNAI1*
04150	mTOR signaling pathway	.006	8.46	*TSC2, PIK3R3*
04650	Natural killer cell mediated cytotoxicity	.006	4.55	*SH2D1A, PIK3R3, ZAP70*
05211	Renal cell carcinoma	.009	5.76	*JUN, PIK3R3*
04640	Hematopoietic cell lineage	.01	4.52	*ITGA6, CD55*
04620	Toll-like receptor signaling pathway	.01	4.42	*JUN, PIK3R3*
04012	ErbB signaling pathway	.01	4.57	*JUN, PIK3R3*

GO indicates Gene Ontology; and KEGG, Kyoto Encyclopedia of Genes and Genomes.

To visualize the classification power of the 62 classification marker genes, we applied an unsupervised 2-dimensional hierarchic clustering algorithm on the expression profiles of the 62 genes across the 535 samples. Remarkably, by only 62 genes, the 6 major ALL subtypes are clearly segregated, with samples of each subtype clustered together (Figure 1A). Interestingly, these markers also separate the unclassified others subtype into 4 major subgroups, with 1 subgroup having similar expression profiles to BCR-ABL. Den Boer et al have reported finding one such subgroup associated with bad prognosis.[40] However, the 535 published samples do not contain prognosis information to allow us to validate Den Boer et al's findings.

Validation of the classifier's performance on a new Chinese children's ALL sample set

Thus far, all the children's ALL samples were derived from white patients. It is not known whether a classifier derived from whites is applicable to a completely different and independent patient population, such as the Chinese ALL children. We therefore collected 100 bone marrow samples from Chinese ALL children, extracted mRNAs, and measured the level of mRNAs using Affymetrix HG-U133A 2.0 microarray. Based on clinical diagnosis, we also categorized the samples into 5 of the 6 subtypes described in Table 1 and traced the 5-year relapse status of all 100 patients.

Due to the storage condition of our samples, we were unable to use flow cytometry to experimentally decide the "hyperdiploid>50" subtype so it was mixed with the others subtype clinically. A total of 44 such mixed samples were combined as a set of "no fusion B-ALL" samples. Among them, our classifier predicted 24 to be hyperdiploid>50, 18 to be others, and 2 mistakenly as *TEL-AML1*. Another misclassification was a *BCR-ABL* sample predicted as others. Estimated from the cross-validation results within the white children's samples (9 samples misclassified between the hyperdiploid>50 and others subtype), an additional 2 misclassified samples would appear if we had distinguished the hyperdiploid >50 from others, thus leading to an estimated approximately 95% overall accuracy for our classifier among the 100 Chinese samples (Table 4 and Figure 1B). Previously the classifiers needed to be retrained every time when encountering a new dataset,[41] however, our gene expression rank-based classifier was directly applied to the new samples without the need to retrain them using the new samples. This made it possible to use the same classifier for any individual future sample, as required by real-world clinical diagnosis. We also found 6 samples of the predicted others subtype in the Chinese patients having expression profiles similar to *BCR-ABL* subgroups based on the 62 classification marker genes (Figure 1B). However, only 2 of the 6 patients relapsed within 5 years after diagnosis and treatment, unlike the 79.2% relapse rate reported by Den Boer et al for a Dutch patient group.[40]

Differentially expressed genes in each ALL subtype

Marker genes, although useful for clinical diagnosis, might not reflect the underlying molecular mechanism of the development of different ALL subtypes. We therefore tried to infer the potential regulatory network that gives rise to the subtype-specific expression patterns for each ALL subtype. As marker genes are only a small subset of differentially expressed genes with the largest differences among subtypes, we first tried to identify the full set of the most significantly differentially expressed genes of each subtype compared with the rest of ALL samples using the 535 white children's samples. For each subtype, the top 50 up-regulated genes and the top 50 down-regulated genes, compared with the rest of ALL samples, were selected based on the *P* value given by RankProd[31] ("Differential gene expression" in "Methods," supplemental Materials), which is a nonparametric statistical method based on a gene permutation model to estimate significance levels of ranks of fold changes between 2 groups of samples.

We found 15.4% of our selected differentially expressed genes were up- or down-regulated in one subtype versus all other subtypes because the genes were oppositely regulated in the rest of the sample versus the normal control. Strictly speaking, such genes are not false positives. However, to facilitate the interpretation of

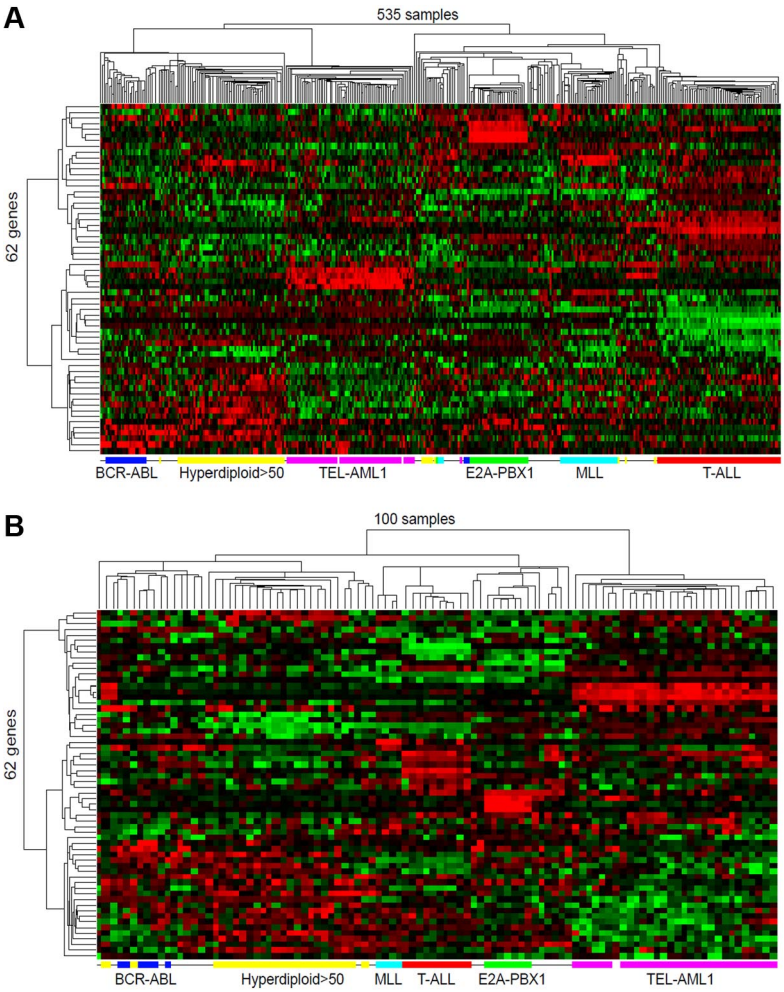

Figure 1. Clustering visualization of the discriminating effects of the marker genes on white and Chinese children's ALL samples. (A) Hierarchical clustering of 535 diagnostic white children's ALL samples (columns) from 3 published datasets using the 62 classification maker genes (rows). (B) Hierarchical clustering of the new 100 diagnostic ALL samples (columns) using the same 62 classification maker genes (rows). Hyperdiploid with more than 50 chromosomes (Hyperdiploid>50) samples in panel B are computationally predicted rather than being experimentally confirmed as in panel A. The expression value for each gene is indicated by color intensity, with red representing high expression and green representing low expression.

differentially expressed genes, we removed such genes using the average expression profile of the 5 normal samples in our new dataset ("Pooling of 5 normal bone marrow samples" in "Methods"), which has not been available in any previous studies. Only the genes that had RankProd *P* value less than .05 compared with the normal control were kept, and the vacancies left by the removed genes in the top 50 list were then filled by the next lower ranked genes. A total of 132 genes were thus replaced. At the end, all of the differentially expressed genes for every subtype had an estimated false discovery rate (given by the "percentage of false positives" ["pfp"] parameter[31]) of 10^{-4} or less compared with the rest of ALL samples.

For all 6 major ALL subgroups, we identified a total of 418 differentially expressed genes, many of which were shared by more than one group. Among these 418 genes, 39 overlapped with the 62 classification markers we identified (supplemental Table 6). The top 50 up-regulated genes of T-cell ALL are enriched in genes of the "T-cell receptor signaling pathways," and the top 50 up-regulated genes of *TEL-AML1* subtype are enriched in "MHC class II protein complex" (supplemental Table 7).

Table 4. Classification accuracies of our 62-gene classifier for each subtype on the 100 Chinese children's ALL samples

Subtype	Accuracy, %	Sensitivity, %	Specificity, %
BCR-ABL	99.0	83.3	100
E2A-PBX1	100	100	100
MLL	100	100	100
No fusion B-ALL*	97.0	95.5	94.6
T-ALL	100	100	100
TEL-AML1	98.0	100	97.2

Total accuracy = 97.0%, which was calculated without classifying the hyperdiploid>50 subtype. Sensitivity and specificity are defined as in Table 2.

ALL indicates acute lymphoblastic leukemia.

*Hyperdiploid>50 and others.

Validation of differentially expressed genes by qPCR

The expression levels of 10 genes on 10 different Chinese samples determined by qPCR generally agree very well with the microarray measurements (R^2 from 0.695-0.946; Figure 2A, supplemental Table 8). We also selected for qPCR validation 8 genes that were determined by RankProd as significantly differentially expressed (pfp $< 10^{-4}$) between 5 *TEL-AML1* samples and 5 *E2A-PBX1* samples from the Chinese children's ALL microarray data. Four of these 8 genes, *CLIC5*, *PCLO*, *PTPRK*, *SOCS2*, were up-regulated in *TEL-AML1* and down-regulated in *E2A-PBX1*, whereas the other 4 genes, *ANKRD15*, *FAT*, *NID2*, *TRIB2*, were the opposite. One-sided *t* test confirmed that all 8 genes were also significantly differentially expressed when their expression levels were quantified by qPCR ($P < .05$, Figure 2B). These differential expression patterns of the 8 genes are also similarly observed among white children's samples (Figure 2B). These results confirmed the high quality of our microarray data as well as the validity of our analysis methods.

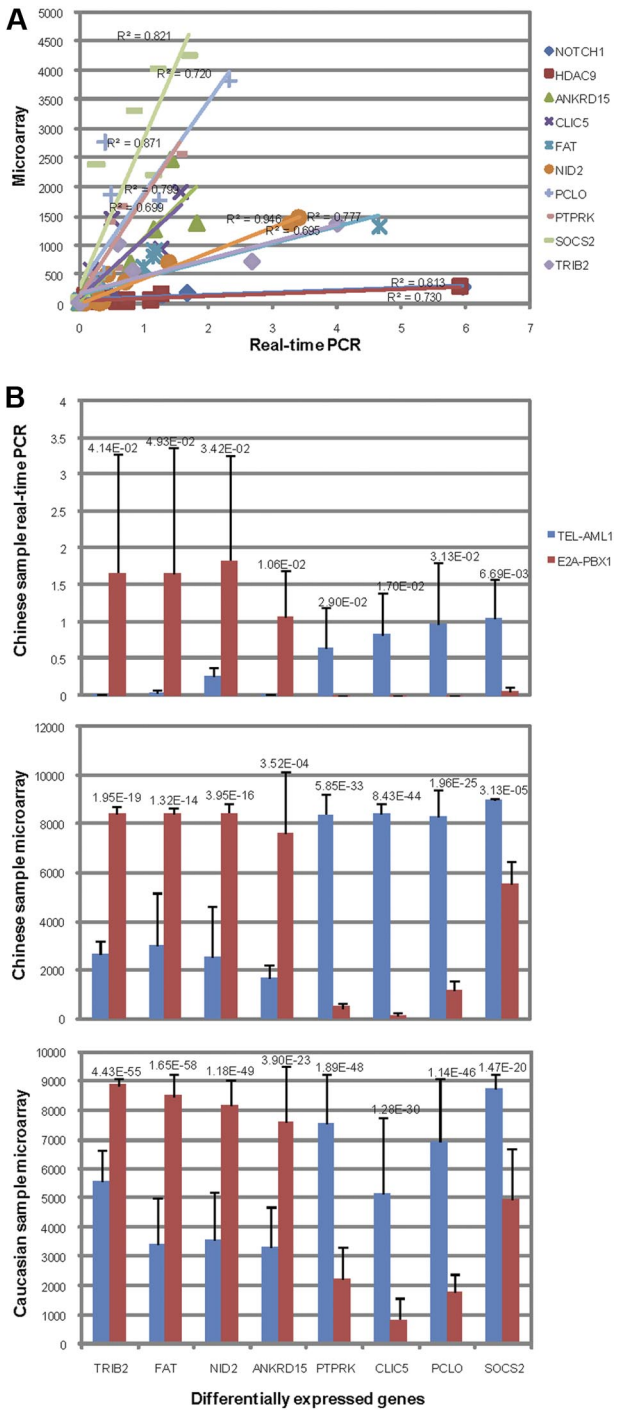

Figure 2. Microarray and qPCR measurements of genes between TEL-AML1 and E2A-PBX1 subtypes. (A) Linear regression of expressions measured by qPCR (x-axis) versus those by microarray (y-axis) of *NOTCH1*, *HDAC9*, *ANKRD15*, *CLIC5*, *FAT*, *NID2*, *PCLO*, *PTPRK*, *SOCS2*, and *TRIB2* in 5 *TEL-AML1* samples and 5 *E2A-PBX1* samples. Linear regression R^2 of each comparison is shown above the corresponding curve. (B) The average expression level of 8 differentially expressed genes in 5 *TEL-AML1* or 5 *E2A-PBX1* samples measured by qPCR (top panel) or by microarray (middle panel) on the same Chinese ALL samples or on white children's samples of the same subtypes (bottom panel). The genes were ordered by the difference between the average expression values of *TEL-AML1* and *E2A-PBX1* samples determined by qPCR (from low to high). The height of each bar represents the average expression level of a gene in a sample group, and the whisker represents the standard deviation. The 1-sided *t* test *P* values between the gene expression values of the 2 ALL subtypes are indicated above the paired bars for each gene.

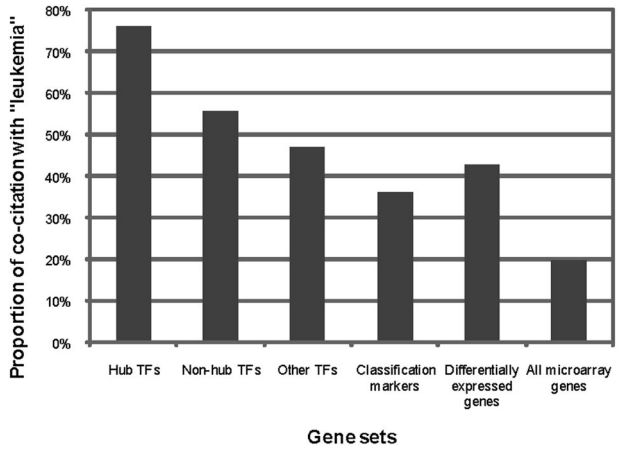

Figure 3. The rate of literature cocitation with the term "leukemia" within different gene groups. The proportion of genes cocited with the term leukemia is the highest among hub TFs in the regulatory networks (21 genes), followed by that of nonhub TFs in our regulatory networks (181 genes), and then those of other human TFs that have binding motifs annotated in the TRANSFAC and JASPAR databases (117 genes), differentially expressed genes (418 genes), and classification markers (62 genes), all of which are significantly higher than the proportion within all genes available on both Affymetrix HG-U133A and HG-U95 microarrays (8178 genes).

Potential regulatory network for each ALL subtype

To find TFs that might be responsible for the subtype-specific differential gene expressions, we first derived potential regulatory interactions between a TF and the differentially expressed genes by the presence of the TF binding motifs in upstream 1-kb sequences of the differentially expressed genes ("Identifying TF-binding motifs for each group of differentially expressed genes" in "Methods"). We also used computationally predicted functional interactions[22] among all differentially expressed genes (including the marker genes) as potential functional relationships among these genes. To identify the interactions relevant to subtype-specific gene expressions, we kept only interactions linking 2 transcriptionally correlated or anticorrelated genes among the 100 Chinese ALL samples (Pearson correlation coefficient $> +0.29$ or < -0.22 corresponding to the top 10% or the bottom 10% of the Pearson correlation coefficients between random gene pairs). The potential interactions among differentially expressed genes and their regulating TFs were visualized as 6 regulatory networks, one for each ALL subtype (supplemental Figure 1). Details of the networks are described in the supplemental Materials (supplemental Figure 1 and supplemental Tables 9-11).

If the TFs were indeed regulators of subtype-specific gene expression, we should expect that they are even more likely to be associated with ALL and its subtypes than the marker genes, and that the more genes in a subtype-specific network a TF targets, the more likely the TF is associated with the disease. We defined hub TFs as the TFs that have out-degree of 4 or higher, corresponding to top 10% of TFs with the highest out-degrees. Among the entire 21 hub TFs in our networks, 76.2% (16 of 21) are known to be related to leukemia (supplemental Table 9), which is much higher than the proportion among nonhub TFs (55.8%, 101 of 181), or that among other human TFs that have binding motifs defined in the TRANSFAC[18,19] and JASPAR[20,21] databases (47.0%, 55 of 117). In contrast, only 35.5% (22 of 62) of the classification marker genes and 42.8% (179 of 418) of the differentially expressed genes are known to be leukemia related (Figure 3). These data suggest that the hub TFs in our predicted regulatory network potentially play very important regulatory roles in the development of ALL.

Discussion

In this study, we first improved gene expression–based ALL subtype classification by compiling and mining a large compendium of samples and using a different marker selection approach, which together led to a high accuracy classifier that can be directly applied to a completely independent sample. Unlike previously reported classifiers,[11-13] ours can take a single sample and make a prediction based solely on the relative expression ranks among the marker genes without consulting the signal distribution of other parallelly processed samples, which is important when dealing with brand new ALL samples. Hoffman et al[13] have obtained a one-time accuracy of 92.6% (without the others group) on a 120 marker gene classifier that has been trained on an old dataset and tested on their new dataset.[13] As cross-validation accuracy in the old dataset has not been tested, it is not known whether there has been a drop of accuracy between their training and test data. In any case, compared with another classifier trained and tested on the old dataset alone by cross-validation, they have shown classifiers generally do not perform as well across datasets as within a dataset. However, our classifier, trained and cross-validated on a panel of 535 white children's samples with very high accuracy, can be applied without any modification to a completely independent set of Chinese patient samples to also achieve very high accuracy. This has been unprecedented thus far in the quest of a truly practical pediatric ALL classifier.

Testing on such an independent sample set also indicates that the second step of selecting the most frequently found markers from 10 different cross-validation groups is necessary to avoid model overfitting. A single SVM classifier trained on all data together could have very high cross-validation accuracy within the dataset, but performed badly on our independent new dataset (data not shown).

We also computationally predicted the potential regulatory networks for the 6 major subtypes and validated the ALL regulatory roles of the predicted transcription regulators through literature mining. The networks provide hints for potential regulatory mechanisms and for new perspectives on clinical treatment of pediatric ALL, which await further functional assays to confirm.

Acknowledgments

This work was supported by National High Technology Research and Development (863) Program-BiotTec Project no. 2006AA02A405 and no. 2006AA02Z4Z2 to Y.H., S.B., and H.Z. and grants from the China National Science Foundation (grant nos. 30890033, 30588001, and 30620120433) and Chinese Ministry of Science and Technology (grant no. 2006CB910700) and funds from the Chinese Academy of Sciences (Bai Ren and KSCX1-YW-R-40) to J.D.J.H. The RNA quality control and microarray hybridization were done by National Engineering Center for Biochip at Shanghai.

Authorship

Contribution: Z.L., M.W., C.G., L.S., and R.Z. collected Chinese ALL samples and performed experimental analysis; W.Z., S.Z., N.Q., and H.X. performed computational analysis; Y.H., S.B., H.Z., and J.-D.J.H., designed and guided the study; and W.Z., H.Z., and J.-D.J.H. wrote the paper.

Conflict-of-interest disclosure: The 62 marker genes and the SVM classifier described here have been submitted for patent (application no. 2009010 82657.6) by all of the authors.

H.Z. represents the Beijing Children's Hospital Leukemia Study Group.

A complete list of the Beijing Children's Hospital Leukemia Study Group participants can be found in the supplemental Appendix.

The current affiliation for H.X. is Department of Genetics, Yale School of Medicine, New Haven, CT.

Correspondence: Jing-Dong J. Han, Chinese Academy of Sciences Key Laboratory of Molecular and Developmental Biology, Center for Molecular Systems Biology, Datun Road, Beijing, 100101, China; e-mail: jdhan@genetics.ac.cn; or Huyong Zheng, Beijing Children's Hospital of Capital Medical University, 56 Nan Lishi Rd, Beijing, China 100045; e-mail: zhenghuyong@vip.sina.com; or Shilai Bao, Center for Molecular Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Datun Road, Beijing, 100101, China; e-mail: slbao@genetics.ac.cn.

References

1. Swensen AR, Ross JA, Severson RK, Pollock BH, Robison LL. The age peak in childhood acute lymphoblastic leukemia: exploring the potential relationship with socioeconomic status. *Cancer.* 1997;79(10):2045-2051.
2. Aricò M, Valsecchi MG, Camitta B, et al. Outcome of treatment in children with Philadelphia chromosome-positive acute lymphoblastic leukemia. *N Engl J Med.* 2000;342(14):998-1006.
3. Biondi A, Cimino G, Pieters R, Pui CH. Biological and therapeutic aspects of infant leukemia. *Blood.* 2000;96(1):24-33.
4. Heerema NA, Sather HN, Ge J, et al. Cytogenetic studies of infant acute lymphoblastic leukemia: poor prognosis of infants with t(4;11): a report of the Children's Cancer Group. *Leukemia.* 1999; 13(5):679-686.
5. Hunger SP. Chromosomal translocations involving the E2A gene in acute lymphoblastic leukemia: clinical features and molecular pathogenesis. *Blood.* 1996;87(4):1211-1224.
6. Pui CH, Evans WE. Acute lymphoblastic leukemia. *N Engl J Med.* 1998;339(9):605-615.
7. Pui CH, Frankel LS, Carroll AJ, et al. Clinical characteristics and treatment outcome of childhood acute lymphoblastic leukemia with the t(4;11)(q21;q23): a collaborative study of 40 cases. *Blood.* 1991;77(3):440-447.
8. Raimondi SC, Behm FG, Roberson PK, et al. Cytogenetics of pre-B-cell acute lymphoblastic leukemia with emphasis on prognostic implications of the t(1;19). *J Clin Oncol.* 1990;8(8):1380-1388.
9. Silverman LB, Gelber RD, Dalton VK, et al. Improved outcome for children with acute lymphoblastic leukemia: results of Dana-Farber Consortium Protocol 91-01. *Blood.* 2001;97(5):1211-1218.
10. Schrappe M, Reiter A, Ludwig WD, et al. Improved outcome in childhood acute lymphoblastic leukemia despite reduced use of anthracyclines and cranial radiotherapy: results of trial ALL-BFM 90: German-Austrian-Swiss ALL-BFM Study Group. *Blood.* 2000;95(11):3310-3322.
11. Yeoh EJ, Ross ME, Shurtleff SA, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell.* 2002; 1(2):133-143.
12. Ross ME, Zhou X, Song G, et al. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood.* 2003;102(8):2951-2959.
13. Hoffmann K, Firth MJ, Beesley AH, de Klerk NH, Kees UR. Translating microarray data for diagnostic testing in childhood leukaemia. *BMC Cancer.* 2006;6:229.
14. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology: The Gene Ontology Consortium. *Nat Genet.* 2000; 25(1):25-29.
15. Open Biomedical Ontologies. The Gene Ontology. http://www.geneontology.org/. Accessed on February, 21, 2008.
16. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000; 28(1):27-30.
17. Kanehisa Laboratory, Bioinformatics Center, Institute for Chemical Research, Kyoto University. KEGG: Kyoto Encyclopedia of Genes and Genomes. http://www.genome.jp/kegg. Accessed on May 10, 2007.
18. Matys V, Fricke E, Geffers R, et al. TRANSFAC:

transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 2003;31(1):374-378.

19. BIOBASE Biological Databases. Transfac. http://www.biobase-international.com/pages/index.php?id=transfac. Accessed March 31, 2006.
20. Bryne JC, Valen E, Tang MH, et al. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* 2008;36(database issue):D102-D106.
21. University of Copenhagen. JASPAR Database. http://jaspar.cgb.ki.se. Accessed November 19, 2008.
22. Xia K, Dong D, Han JD. IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. *BMC Bioinformatics.* 2006;7:508.
23. HanLab. IntNetBD. http://hanlab.genetics.ac.cn/sys. Accessed January 5, 2009.
24. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug-target network. *Nat Biotechnol.* 2007;25(10):1119-1126.
25. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* 2003;4(2):249-264.
26. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 2004;5(10):R80.
27. National Center for Biotechnology Information. Gene Expression Omnibus (GEO). http://www.ncbi.nlm.nih.gov/geo. Accessed August 26, 2009.
28. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning.* 2002;46:389-422.
29. Xia K, Xue H, Dong D, et al. Identification of the proliferation/differentiation switch in the cellular network of multicellular organisms. *PLoS Comput Biol.* 2006;2(11):e145.
30. http://www.bch.com.cn/xy/BCH_ALL_microarray_data.rar. Accessed August 26, 2009.
31. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics.* 2006;22(22):2825-2827.
32. Smith AD, Sumazin P, Xuan Z, Zhang MQ. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci U S A.* 2006;103(16):6275-6280.
33. Karolchik D, Hinrichs AS, Furey TS, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 2004;32(database issue):D493-D496.
34. Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco, CA: Morgan Kaufmann; 1999.
35. Schölkopf B, Burges CJC, Smola AJ. *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA: MIT Press; 1999.
36. ACM Digital Library. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM; 1999.
37. Dasarathy BV. *Nearest Neighbor (NN) norms: NN Pattern Classification Techniques*. Los Alamitos, CA: IEEE Computer Society Press; 1991.
38. Raedt Ld, Flach P. Machine Learning: ECML 2001: 12th European Conference on Machine Learning, Freiburg, Germany, September 5-7, 2001: *Proceedings*. New York, NY: Springer-Verlag; 2001.
39. Zhou X, Tuck DP. MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. *Bioinformatics.* 2007;23(9):1106-1114.
40. Den Boer ML, van Slegtenhorst M, De Menezes RX, et al. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. *Lancet Oncol.* 2009;10(2):125-134.
41. Kohlmann A, Schoch C, Schnittger S, et al. Pediatric acute lymphoblastic leukemia (ALL) gene expression signatures classify an independent cohort of adult ALL patients. *Leukemia.* 2004;18(1):63-71.