Journal of Southeast University (English Edition) Vol. 22, No. 2, pp. 248 – 255 June 2006 ISSN 1003—7985

Meta-analysis gene lists about subtypes of leukemia based on gene expression data

Yang Xinan Sun Xiao Lu Zuhong

(State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China)

Abstract: To screen for molecular signatures that are commonly dysregulated in subtypes of a certain cancer, a novel meta-analysis is designed to perform rank score (RS) on lists of genes that are derived from different studies. RS is a promising way to detect signatures across platforms when integrating with one vs. all (OVA) or one vs. one (OVO) schemes of comparison. Among six published microarray expression datasets on acute leukemia, the biological signals hereafter provide stronger clustering support than systematic differences among microarray platforms. Moreover, the pediatric BCR_ABL specific genes can be used to correctly discriminate independent adult BCR_ABL cases. The obtained results redound to discover, validate and treat the subtypes from microarray gene expression profiles of cancer, which have been plentifully researched, such as leukemia.

Key words: oligonucleotide microarray; meta-analysis; rank score; leukemia

Leukemia is a malignant cancer that originates from precursor blood cells in the bone marrow. Treatment for leukemia is complex and highly dependent on the special disease type. Gene expression profiling combined with advanced bioinformatics has begun to disclose the genes enacting specific biochemical events leading to special types of leukemia[1], as well as the changes in gene expression caused by specific treatments[2]. Diagnosis of leukemia involving multiple categories is generally more difficult than in the case of two categories. It is also more difficult to differentiate subtypes of cancer with similar biomolecular pictures than those with distinctive appearances. Although leukemia is among the best-studied cancers on microarray, there are special subtypes which cannot be classified with high accuracy, such as ALL cases bearing the t(9;22)(BCR_ABL) translocation[3].

First, we investigate how to compare gene lists obtained from different studies. Given the increased availability of gene expression profiles on leukemia, there is a new chance and challenge to reuse, merge and compare these data to calibrate knowledge of leukemia. It has been widely accepted that meta-analysis is a way to identify differentially expressed genes among tumor types across microarray platforms[4-6]. The question as to whether the results of gene expression measurements obtained by different platforms can be compared has been addressed[7-8]. In contrast to previous research, which integrated expression measurement of gene profiles from individual studies, we address the problem of how to combine the statistics of all measured genes obtained from different studies for further analysis of classification.

Secondly, we investigate how to detect the important genes with regard to a certain subtype in the leukemia community with multiple phenotypes. One can yield a set of genes by using a one vs. all (OVA) scheme for subtype classification[9-10]. But what does it mean when some of these marker-genes of subtype A also discriminate other subtype B from the rest? Overlaps in the gene list are to be expected because the signatures of a distinctive group besides A and B will strongly affect the cases labeled as "not A" as well as "not B" by the OVA scheme. In the leukemia world, T-ALL is the extruding subtype. We use two ways to eliminate these overlaps, since we just want sets of genes that are specific for the given subtype. One method is to find intersections between the OVA comparisons and discard the inter-type overlapping genes. Another way, which is more promising but seldom used in the context of microarray data analysis before, is to really perform one vs. one (OVO) analysis for subtypes and only keep inner-type intersections of genes.

Compared to OVA comparisons, where the "all other subtypes" may combine on an unbalanced expressed level of subtypes, OVO comparison specifies

Received 2005-11-21.

Foundation items: The National High Technology Research and Development Program of China (863 Program) (No. 2002AA231071), the Natural Science Foundation of Jiangsu Province (No. BK2002057), the Sandwich Scholarship by German Academic Exchange Service.

Biographies: Yang Xinan (1971—), female, graduate; Lu Zuhong (corresponding author), male, doctor, professor, zhlu@seu.edu.cn.

the detection of differential expression for given subtypes of leukemia. We apply our method to six microarray studies involving one on leukemia drug response, one on subtypes of adult leukemia, and four on subtypes of pediatric leukemia. The result shows that the OVA-RS or the OVO-RS scheme yields more sensitive and reliable clusters, based on biological features rather than on systematic differences among different Affymetrix microarrays. We also validated those marker-genes for subtypes of leukemia on an independent study.

1 Materials and Methods

1.1 Data collection and processing

We collect microarray data produced on Affymetrix arrays published recently. The datasets used in this study are all publicly available. Of these, we use five[3, 11 - 14] for identifier detection and one for identifier validation[15].

For all these datasets, profiles are of mRNA samples. The preprocession, such as background correction, normalization, summarization and quality assessment on all those microarray datasets, are independently based on Affymetrix platform preprocessing protocol. We normalized the raw probe expression values by variance stabilization and subsequently summarized the probe values into one probe set/gene expression value using median polish[16]. Thereafter, our expression data have a generalized logarithmic scale, meaning that additive differences correspond to fold-changes on this scale.

1.2 Initial data analysis

We assign leukemia samples to groups according to their diagnostic subtypes given by the original authors. Then we exclude the subtypes which contain less than eight samples from one study. A fold changes (FC) test is conducted as two-sided for differential expression analysis. Let the vector $\boldsymbol{y}^d \in \{A^d, B^d, C^d, \ldots\}$ contain the subtype labels of patients and matrix $\boldsymbol{E}^d = (e_{ij}^d)$ contain the actual expression data in study d. The expression data has been normalized as described above and has a logarithmic scale. Note that on this scale, additive differences correspond to FC in actual molecule abundance.

Given a study d and a pair of diagnostic groups, we compare the expression levels of patients in group A to those in group B for each gene. To this end, we calculate the differences between two groups:

$$f_i^d = e_i(\bar{A}^d) - e_i(\bar{B}^d)$$

where $e_i(\bar{A}^d)$ and $e_i(\bar{B}^d)$ denote the average expression of gene i in group A and B, respectively. This vector of fold-change statistics f_i^d is afterwards called as entity in this paper.

In order to compare the entity f_i^d to FC statistics expected by chance, we observe N permutations of the class labels and compute empirical p-values[17]. To this aim, we randomly shuffle the class labels $\boldsymbol{y}^d$ in each study $N(=1\,000)$ times. From the randomized data, we calculate sets of entities in the same way as for the original data. Thus we get N vectors of randomized values of FC f_i^n, $n = 1, 2, \ldots, N$ for each entity, from which we derive empirical p-values:

$$p_i^d = 1 - N^{-1} \mid \{f_i^d < f_i^n\} \mid$$

where $\mid \cdot \mid$ denotes cardinality.

This yields two kinds of matrices for each study d, one of FC statistics $\boldsymbol{F}^d = (f'_{ik})$ and another of empirical p-values $\boldsymbol{P}^d = (p'_{ik})$. The columns $k' \in 1, 2, \ldots, K^d$ hold the K^d pairwise entities between groups, and rows correspond to transcripts in study d.

1.3 Rank scores

Microarray data from different platforms are thought to be not directly comparable, because they often use distinct reference samples and different protocols. Even for those different studies using the same platform and analysis technique, it is still difficult to compare their values directly. The idea of "rank score" is driven from the "median rank score" for measurement of gene expression across multiple studies[7]. In this paper, we apply it to gene entities before performing meta-analysis.

First, we reduce the entities in all studies to those transcripts, whose expressions have been measured on all microarrays in the five studies. The matching is done using the spreadsheet given by Affymetrix online support. A number of 8 620 "best match" transcripts are common to all studies.

To render statistics that come from different platforms comparable to each other, we build a numerical scale from all entities and all common transcripts. Having calculated all the possible entities in five studies, we yield the integrated FC matrix $\boldsymbol{X} = (f_{ik}), k = 1, 2, \ldots, \sum_d K^d$ combining of each $\boldsymbol{F}^d = (f'_{ik})$, according to the common transcripts. We separately sort all entities by their transcripts' FC values. Then for each rank, we take the mean value over all entities' values at each rank regardless of the transcripts annotated there. This yields an entity of reference values r. Next, for each o-

riginal entity, we map them to a fixed numerical scale to generate a matrix of rank score (RS) $\boldsymbol{X}^* = (x_{ik}^*)$ according to their ranks:

$$x_{\cdot k}^* = \overline{\text{sort}(f_{\cdot k})}[\text{rank}(f_{\cdot k})]$$

where $\overline{\text{sort}(f_{\cdot k})}$ takes the mean value across all the samples for each rank. Thereafter, all entities from five data sets contain comparable statistics on the same numerical scale.

1.4 Clustering subtypes of leukemia

After making entities and putting them on the same scale, we investigate how to identify subtypes of leukemia to detect the genes that are specific for the given subtype. Using the OVA method, we know which signature belongs to which subtype and then classify the across-study-entities according to their subtypes.

In contrast, we need to identify the cluster of subtypes for those OVO pairwise comparisons. To this end, we cluster all entities, and then identify subtypes represented by these clusters such that the most common subtypes belong to it. Therefore, we adjust the RS: Treat non-significantly changed values similar to unchanged values in each entity to filter out noise. First, we take the absolute RS because that A^{d1} vs. B^{d1} from one study should be clustered with A^{d2} vs. B^{d2} from another study. Secondly, the FC test can detect genes of interest, but it has the obvious disadvantage in that it does not provide an estimate of significance for the observed changes and thus the necessary cutoff values[18]. Therefore we filter the non-significant changed genes in each entity. Given a significant threshold T, we define the matrix of signatures $\boldsymbol{S}$ as

$$s_{ik} = \begin{cases} 1 & p_{ik} < T,\ f_{ik} \geqslant 0 \\ -1 & p_{ik} < T,\ f_{ik} < 0 \\ 0 & p_{ik} \geqslant T \end{cases}$$

where p_{ik} is the element of the integrated p-value matrix $\boldsymbol{P} = (p_{ik}), k = 1, 2, \ldots, \sum_d K^d$, combining of $\boldsymbol{P}^d$, and f_{ik} is the element of the integrated FC-value matrix $\boldsymbol{X}$. Thus one element of $\boldsymbol{S}$ is a signature that the corresponding transcript significantly changed in the corresponding entity. Thereafter, every element of the RS is multiplied by the corresponding element in $\boldsymbol{S}$ as

$$y_{ik} = x_{ik}^* \ s_{ik}$$

In this way we filter the RS into an adjusted RS (ARS): $\boldsymbol{Y} = (yi_k)$.

Then bottom-up hierarchical cluster analysis is performed to find clusters of OVO entities based on the matrix $\boldsymbol{Y}$. Note that each entity (column) in $\boldsymbol{Y}$ is a pairwise entity of one subtype vs. another. Euclidean distance between the columns of $\boldsymbol{Y}$ is taken for clustering using complete linkage. Note that we cluster entities that refer to pairs of subtypes. In the event that the majority of entities in one cluster involved a special subtype of leukemia, we argue that this cluster represents the molecular characteristics of such a subtype, and thus label that cluster accordingly. Therefore, we can assign each entity a label of subtype. To this end, we set the desired number of clusters $c = 16$ to the number of subtypes of interest. Then we identify subtypes represented by these clusters such that the most common subtype belongs to it. For example, if a cluster consists of nine entities which referred to T-ALL vs. another subtype, we will label this cluster as "T-ALL".

1.5 Meta-signatures of subtypes of children leukemia

We now seek to identify the meta-signatures that characterize certain subtypes of leukemia across multiple platforms or studies. The first type of analysis is performed on OVO entities. We pick up the common genes which are significantly up- or down-expressed across all the entities which are clustered to represent certain subtypes. For example, all 52 transcripts (see Tab. 1) are dysregulated for the cluster ID. 2 as shown in Fig. 1, representing the entities of MLL vs. another subtype. In Fig. 1, the cluster tree is based on ARS filtered by p-value ($p < 0.01$). The label of one entity is formed as "x: A/B - - y", where $x = 1, 2, \ldots, 5$ means the ID number of datasets in Tab. 2; A/B refers to the entities of group A vs. group B; and $y = 1, 2, \ldots, 16$ denotes the resulting ID number of clustering. The result is interesting that the biological signal is stronger than the systematic differences between platforms. For example, the entities of T-ALL vs. other leukemia subtypes from different studies are clustered together into one cluster with ID No. 5, MLL into ID No. 2, E2A-BPX1 into ID No. 3, BCR_ABL into ID No. 4, and FAB-M7 into ID No. 14, etc. On the bottom is the number of signatures in each entity, which survives the threshold $p < 0.01$.

In contrast, we have to find intersections between the signature of subtypes and discard them when using the OVA scheme. In this study, we focus on the nine subtypes of ALL. To this end, we assign the marker-genes of subtypes as the consistently significant ($p < 0.01$) genes in OVA entities respected to the given subtypes of ALL. Then for each set of subtype-marker-genes, we remove those appearing in the other eight sets of genes to make sure they are "subtype-specific".

Tab. 1 52 children OVO MLL specific genes

Probe	Symbol	Chromo-some	Up/down	Probe	Symbol	Chromo-some	Up/down	Probe	Symbol	Chromo-some	Up/down
1389 _ at	MME	3	–	36536 _ at	SCHIP1	3	–	40396 _ at	P2RX5	17	+
1914 _ at	CCNA1	13	+	36650 _ at	CCND2	12	–	40451 _ at	POLE	12	–
2036 _ s _ at	CD44	11	+	36777 _ at	KLRK1	12	+	40493 _ at	CD44	11	+
2062 _ at	IGFBP7	4	+	36937 _ s _ at	PDLIM1	10	–	40518 _ at	PTPRC	1	+
266 _ s _ at	CD24	6	–	37043 _ at	E2F2	1	–	40520 g _ at	PTPRC	1	+
307 _ at	ALOX5	10	–	37421 _ f _ at	HLA – F	6	–	40522 _ at	GLUL	1	+
31472 _ s _ at	CD44	11	+	37479 _ at	CD72	9	+	40729 _ s _ at	NCR3	6	–
32193 _ at	PLXNC1	12	+	37809 _ at	HOXA9	7	+	40763 _ at	MEIS1	2	+
32207 _ at	MPP1	X	–	37864 _ s _ at	IGHG3	14	–	40797 _ at	ADAM10	15	+
32607 _ at	BASP1	5	+	38194 _ s _ at	IGKC	2	–	40913 _ at	ATP2B4	1	–
33412 _ at	GGA1	22	+	38287 _ at	PSMB9	6	–	41266 _ at	ITGA6	2	–
33705 _ at	PDE4B	1	–	38291 _ at	PENK	8	+	41448 _ at	EVX1	7	+
34168 _ at	DNTT	10	–	38391 _ at	CAPG	2	+	41470 _ at	PROM1	4	+
34210 _ at	CD52	1	–	38413 _ at	DAD1	14	+	41710 _ at	LOC54103	7	–
34306 _ at	MBNL1	3	+	39318 _ at	TCL1A	14	–	657 _ at	PCDHGC3	5	+
34785 _ at	THRAP2	12	+	39327 _ at	D2S448	2	–	769 _ s _ at	ANXA2	15	+
35663 _ at	NPTX2	7	+	39338 _ at	S100A10	1	+	794 _ at	PTPN6	12	+
36239 _ at	POU2AF1	11	–								

Note: Those genes commonly dysregulated in the 13 entities derived from the expression profiles of three pediatric leukemia studies. The sign" + "means the fold change statistics is higher in MLL leukemia samples than in other samples, while the opposite cases are signed as " – ".

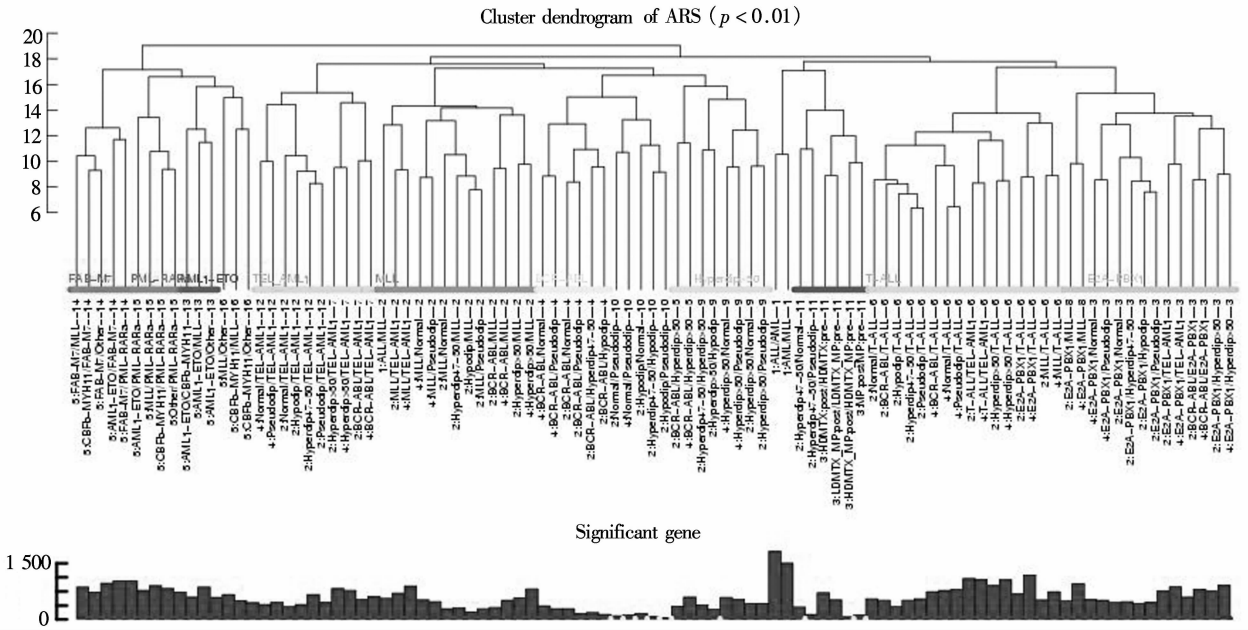

Fig. 1 95 OVO entities on sub-disease from 5 leukemia studies are clustered into 16 clusters

2 Results

We catalogue information on 741 pediatric leukemia microarray samples from five published studies for marker-gene detection, based on Affymetrix technology. Each study examines subtypes or drug-response of the pediatric leukemia cases, and each one is respectively preprocessed using the package compdiagTools. Among these samples, 728 samples meet the requirement that they belong to a subtype consisting of more than eight samples in that study. One adult leukemia[15] research is used to check whether the signature derived from pediatric is validated for adult data preprocessed for additive scale by the package vsn. Both compdiagTools and vsn include the tools for variance stabilization and calibration for microarray data.

2.1 Data and entities for gene detection

The first data set[11] reports the assumption that MLL (n = 20, Myeloid/lymphoid or Mixed Lineage Leukemia) should be distinguished from both AML (n =28) and other ALL(n =24), that challenge the traditional two classes of acute leukemia. The second data-

set is a study on pediatric ALL[12]. Here, 327 leukemia samples fall into nine subtypes of B-precursor ALL and T-cell lineage ALL. The third study aims at detecting treatment-specific changes in gene expression upon four different treatments[13]. The gene expression profiles of bone marrow leukemia cells before treatment and one day after treatment are examined. Another dataset chooses 132 ALL tissue samples from the above 327 samples[12]. It consists of all the 10 subtypes of ALL[3]. The last study identifies expression signature for 130 pediatric patients with seven subtypes of AML[14]. The more details are given in Tab. 2.

Tab. 2 Key characteristics of six independent data

ID	Paper	Platform
1	Ref. [11]	HG-U95a
2	Ref. [12]	HG-U95av2
3	Ref. [13]	HG-U95av2
4	Ref. [3]	HG-U133a
5	Ref. [14]	HG-U133a
6	Ref. [15]	HG-U95av2

After raw profiling has been normalized on additive scale, each gene is assessed for differential expression with an FC equivalent test using twilight package in Bioconductor[17]. The FC test and the t-test yield the same cluster results and similar multidimensional scaling plots. We calculate all the possible comparisons from four studies except the study of drug response[13], where only the diagnostically meaningful entities of pre- vs. post-treatment are considered. There are 95 OVO entities and 31 OVA entities.

We filter the entity without significant ($p < 0.01$) signature. All 95 OVO and OVA entities have at least one signature. The corresponding combined matrices $\boldsymbol{X}$ and $\boldsymbol{P}$ are yielded by picking out the common best-matched 8 620 transcripts from all studies.

2.2 Signature across studies vs. that across phenotypes

It is obvious that the RS can cluster the subtypes of leukemia instead of the source of studies. These RS map the FC values into the same numeric scale without losing the orders of expressed genes. Applying ARS, the biological signatures are even stronger due to filtering the noise. Performing principal coordinate analysis, we can separate the major types of leukemia well. Using either the OVA or the OVO scheme, AML, T- and B-cell ALL can be separated clearly in two-dimensional spaces, so do other important subtypes of AML and B-cell ALL.

Moreover, the multiple subtypes of leukemia can be clearly separated after adopting ARS and performing hierarchical clustering on Euclidean distance of OVO or OVA entities. Fig. 1 demonstrates the ARS result of OVO entities given a significant threshold of $p < 0.01$. Most subtypes of leukemia can be identified as one cluster, no matter what the study of the entities. Salient, four subtypes are so distinguishing that all OVO entities referring to it are clustered together. These are T-ALL, E2A-PBX1 of ALL, and FAB-M7, PML-RARa of AML. The signatures in the cluster of BCR_ABL is adjacent to that of ALL chromosome abnormalities using the OVO scheme, and to that of hyperdiploid > 50 adopting the OVA scheme.

We list these marker-genes in Tab. 1 for MLL and Tab. 3 for BCR_ABL subtype. In Tab. 3, those genes commonly dysregulate in two entities derived from the expression profiles of two pediatric leukemia studies. The sign "+" means the fold change statistics is higher in BCR_ABL leukemia samples than in other samples, while the opposite cases are signed as "–". Since rearrangements of the MLL gene occur in both acute lymphoblastic and acute myeloid leukemias (ALL, AML), it is promising to mine the potential MLL marker-genes from the datasets of both types. We will discuss these marker-genes in the next section.

Tab. 3 23 children OVA BCR_ABL specific genes

Probe	Symbol	Chromosome	Up/down
1983_at	CCND2	12	+
33362_at*	CDC42EP3	2	+
33924_at	RAB6IP1	11	+
34237_at*	HBS1L	6	+
34644_at	B2M	15	+
34877_at	KIAA1579	1	+
36035_at	GPAA1	8	+
36138_at	CAPNS1	19	+
36660_at	RAB11A	15	+
37112_at	C6orf32	6	+
37347_at	CKS1B	1	–
37652_at	CABIN1	22	+
37762_at	EMP1	12	+
38032_at	SV2A	1	+
38077_at	COL6A3	2	+
38312_at*	OLFML2A	9	+
39753_at	ITGA5	12	+
40051_at	TRAM2	6	+
40196_at	CTDSPL	3	+
40202_at	KLF9	9	+
40504_at*	PON2	7	+
41753_at	ACTN4	19	+
41872_at*	DFNA5	7	+

Note: The probes signed with * are detected in the OVO method as well.

2.3 Validation on the marker-genes of subtypes of child leukemia

Some subtypes of ALL are so widely distinct from

other subtypes that more than 500 genes significantly differential in the OVA way in both studies[3,12]. These subtypes are T-ALL, E2A-PBX1, TEL-AML1, hyperdiploid >50.

To assess the validation of these subtype marker-genes we first reuse the dataset of Ross[3] with phenotype information in details. Six OVA or OVO sets of ALL-subtype-specific genes are used in the linear SVM-based supervised learning algorithm for training classifier. It is done in a randomly selected training set that consists of three fourths of the total cases (100 cases) as Ross et al. did in their paper. The result choruses the diagnostic subtypes of pediatric ALL (for details see Tab. 4). In Tab. 4, S_1 represents sensitivity, S_2 represents specificity. We adopt the discriminating ($p < 0.05$) genes and supervise the learning algorithm developed from 132 ALL cases. The columns labeled as "original result" are the results reported by Ross et al. The last column of "OVO ARS" refers to the ID number identified for known subtypes as shown in Fig. 1. "#genes1" represents the number of marker-genes being inner-type shared. "#genes2" represents the number of inner-type shared genes without inter-type overlapping. We find that:

- Three subtypes of ALL, namely T-ALL, E2A-PBX1, and TEL-AML1, have the strongest gene-expression signature. A prediction accuracy of 99% to 100% can be achieved using both the OVA and the OVO schemes. While the OVO scheme identifies fewer genes than OVA does.
- But for the subtypes having weaker gene-expression signatures such as BCR_ABL and hyperdiploid >50, fewer OVA marker-genes give higher accuracy than OVO marker-genes do.
- Interestingly, the 52 OVO marker-genes for MLL arrangement correctly predict the 20 cases from the study of Ross et al[3]. But in the OVA scheme, no genes can be identified.

Tab. 4 Prediction accuracies in percent of ALL subtypes

Subtypes	Original result		OVO ARS				OVA ARS			
	S_1/%	S_2/%	S_1/%	S_2/%	Cluster	ID	#genes 1	S_1/%	S_2/%	#genes 2
T-ALL	100	100	100	100	6	135	100	100	1047	428
E2A-PBX1	100	100	100	100	3 and 8	99	100	100	598	119
TEL-AML1	100	100	100	99	7 and 12	74	100	100	634	136
BCR_ABL	75	100	85	100	4	57	92	99	228	23
MLL	100	100	100	100	2	52	—	—	243	0
Hyperdiploid >50	100	100	78	96	9	85	94	99	587	134

Next, we assess our leukemia specific signature on an independent dataset[15]. This data is different from all the above data in that they are from adult cases. We are able to obtain 100% diagnostic accuracy in distinguishing B-cell from T-cell for the 128 adult leukemia patients, using the genes identified from pediatric ALL cases. Moreover, for the adult BCR_ABL subtype diagnosed in a molecular biological way, only 11 cases are mis-classified ($S_1 = 93.4\%$, $S_2 = 86.5\%$) using the OVO marker-genes ($n = 57$) identified from pediatric cases in two studies. In contrast, using OVA marker-genes ($n = 23$), 20 cases are mis-classified ($S_1 = 89.0\%$, $S_2 = 73.0\%$). BCR_ABL is the complex karyotype of ALL. Our results reveals a new way for elaborate diagnosis.

3 Discussion

Our study provides a simple, scalable design of meta-analysis to evaluate, integrate and cluster the standard test results of multiple datasets. First, RS substitutes all statistical values with one reference numerical scale according to their ranks to make all entities comparable. This presumably changes variance for some genes but it most notably fits all entities to the same numerical scale while keeping the level of differential expression in each entity. Secondly, it is diagnostically helpful to identify the pure subtype-specific marker-genes, though genes are thought to be coregulated. Rifkin et al.[10] showed that the OVA scheme, in combination with the support vector machine, gave the most accurate method by a significant margin. Our experiment on ALL data suggests that the OVA scheme is suitable for subtypes with weak signatures, while the OVO scheme is more promising for subtypes with strong signatures. Our results suggest that the tradeoff between OVA and OVO should be carefully counted in multiple classification.

Leukemia is among the best-studied cancer based on microarrays. However, two subtypes t(9; 22) (BCR_ABL fusion) and MLL-rearrangement in ALL, carry an unfavorable prognosis and still contain diagnostic errors[2]. By our method, the signatures of these two subtypes are significantly distinguishing across platforms and studies. The 23 OVO MLL marker-genes are detected from three independent studies including both AML and ALL cases. Among them, 11 are

involved in response to biotic stimulus (hypergeometric $p = 10^{-9}$), namely CD24, DNTT, POU2AF1, KLRK1, PDLIM1, CD72, IGHG3, PSMB9, CAPG, and NCR3. Moreover, CD24 was reported as one of marker-genes for adult MLL rearrangement[19]. Terminal deoxynucleotidyl transferase (DNTT) is a unique intranuclear DNA polymerase that catalyzes the template-independent addition of deoxynucleotides to the 3'-hydroxyl terminus of oligonucleotide primers. These results support that down-regulated DNTT might be associated with MLL gene rearrangement[20]. POU2AF1 provides structural and functional specificity in the regulation of immunoglobulin transcription, and is proposed as a potential proto-oncogene for MLL-AF4[21]. Interestingly, PTPRC (protein tyrosine phosphatase, receptor type, C) is involved in T-cell selection[22], and CD44 is the regulator of T-cell activation[23]. Here we again show that MLL fusions might also generate a distinct genetic subtype of T-lineage ALL[24]. Taken together, the identified MLL marker-genes reveal new insights into the aberrant transcriptional program MLL leukemias. In addition, two in the 23 OVA BCR_ABL marker-genes are involved in cytokinesis (hypegometric $p = 0.029$), namely CCND2 and CKS1B. To our knowledge, these genes are firstly reported as marker-genes for BCR_ABL and correctly classified this subtype of leukemia (see Tab. 4). Thus our result reveals a possible new way for leukemia classification and marker-gene detection. Further insights into the subtype specific genes might help to understand the abnormal growth of leukemia.

The salient finding is as follows: Our results suggest that the cell type discriminating genes which are identified from the pediatric ALL cases can be used to accurately diagnose adult cases. The analogous result has been reported in Ref. [14] for children and adult AML cases. It promotes more entities on pediatric and adult ALL for identification of class discriminating and related genes with more published data. Our results also suggest a high risk of chromosome abnormalities in BCR_ABL patients group.

Gene expression data can be obtained from arrays containing cDNA clones or oligonucleotides, or other gene-specific PCR products. Although grossly similar, microarray platforms differ in sequence content and measurement methodology and thus produce qualitatively different data. For example, cDNA microarray performs a two-color competitive hybridization that gives the ratio of transcript expression in two samples. In contrast, oligonucleotide chips, such as those provided by Affymetrix, provide an absolute measurement of gene expression in one sample. For simplicity, we only collect recently published microarray data produced by Affymetrix arrays. However, previous evaluations of microarray technologies[25] found strong correlations ($r = 0.8$ to 0.9) among relative gene expression measurements made with different microarray technologies. Therefore it would be more important to collect leukemia data based on more kinds of platforms, e. g. cDNA chips, to see whether the biological signal is always stronger than the systematic differences between multiple platforms.

In summary, our work established the methodology to classify certain cancers based on meta-analysis of gene expression profiling, and demonstrated its effectiveness with application on leukemia studies. It is shown that careful design of meta-analysis can exact and extend clues into leukemia. We believe that these genes and clues identified by ARS will helpfully expand our knowledge of the mechanism of leukemia progression and treatments.

References

[1] Pui C H, Schrappe M, Ribeiro R C, et al. Childhood and adolescent lymphoid and myeloid leukemia [J]. *Hematology*, 2004, **2004**(1): 118 − 145.

[2] Holleman A, Cheok M H, den Boer M L, et al. Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment [J]. *N Engl J Med*, 2004, **351**(6): 533 − 542.

[3] Ross M E, Zhou X, Song G, et al. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling [J]. *Blood*, 2003, **102**(8): 2951 − 2959.

[4] Rhodes D R, Yu J, Shanker K, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression [J]. *Proc Natl Acad Sci USA*, 2004, **101**(25): 9309 − 9314.

[5] Segal E, Friedman N, Koller D, et al. A module map showing conditional activity of expression modules in cancer [J]. *Nature Genet*, 2004, **36**(10): 1090 − 1098.

[6] Grutzmann R, Boriss H, Ammerpohl O, et al. Metaanalysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes [J]. *Oncogene*, 2005, **24**(32): 5079 − 5088.

[7] Toedling J, Spang R. Assessment of five microarray experiments on gene expression profiling of breast cancer [A]. In: *The Seventh Annual International Conference on Research in Computational Molecular Biology* [C]. Berlin, http://citeseer.ist.psu.edu/611350.html. 2003.

[8] Warnat P, Eils R, Brors B. Cross-platform analysis of cancer microarray data improves gene expression based

classification of phenotypes [J]. *BMC Bioinformatics*, 2005, **6**(1): 265.
[9] Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures [J]. *Proc Natl Acad Sci USA*, 2001, **98**(26): 15149 - 15154.
[10] Rifkin R, Mukherjee S, Tamayo P, et al. An analytical method for multi-class molecular cancer classification [J]. *SIAM Reviews*, 2003, **45**(4): 706 - 723.
[11] Armstrong S A, Staunnton J E, Silverman L B, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia [J]. *Nature Genet*, 2002, **30**(1): 41 - 47.
[12] Yeoh E J, Ross M E, Shurtleff S A, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling [J]. *Cancer Cell*, 2002, **1**(2): 133 - 143.
[13] Cheok M H, Yang W, Pui C H, et al. Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells [J]. *Nature Genet*, 2003, **34**(1): 85 - 90.
[14] Ross M E, Mahfouz R, Onciu M, et al. Gene expression profiling of pediatric acute myelogenous leukemia [J]. *Blood*, 2004, **104**(12): 3679 - 3687.
[15] Chiaretti S, Li X, Gentleman R, et al. Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival [J]. *Blood*, 2004, **103**(7): 2771 - 2778.
[16] Huber W, von Heydebreck A, Sueltmann H, et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression [J]. *Bioinformatics*, 2002, **18** (Sup. 1): 96 - 104.
[17] Scheid S, Spang R. Twilight: a bioconductor package for estimating the local false discovery rate [J]. *Bioinformatics*, 2004, **21**(12): 2921 - 2922.
[18] Breitling R, Armengaud P, Amtmann A, et al. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments [J]. *FEBS Letters*, 2004, **573**(1 - 3): 83 - 92.
[19] Schwartz S, Rieder H, Schlgger B, et al. Expression of the human homologue of rat ng2 in adult acute lymphoblastic leukemia: close association with mll rearrangement and a cd10(-)/cd24(-)/cd65s(+)/cd15(+) B-cell phenotype [J]. *Leukemia*, 2003, **17**(8): 1589 - 1595.
[20] Liu L, McGavran L, Lovell M A, et al. Nonpositive terminal deoxynucleotidyl transferase in pediatric precursor blymphoblastic leukemia [J]. *Am J Clin Pathol*, 2004, **121**(6): 810 - 815.
[21] Yuille M A, Galiegue-Zouitina S, Hiorns, et al. Heterogeneity of breakpoints at the transcriptional co-activator gene, bob-1, in lymphoproliferative disease [J]. *Leukemia*, 1996, **10**(9): 1492 - 1496.
[22] Sawa H, Nagashima T, Nagashima K, et al. Clinicopathological and virological analyses of familial human tlymphotropic virus type Ⅰ—associated polyneuropathy [J]. *J Neurovirol*, 2005, **11**(2): 199 - 207.
[23] Koga H, Imada K, Ueda M, et al. Identification of differentially exp, ressed molecules in adult t-cell leukemia cells proliferating in vivo [J]. *Cancer Sci*, 2004, **95**(5): 411 - 417.
[24] Ferrando A A, Armstrong S A, Neuberg D S, et al. Gene expression signatures in MLL-rearranged tlineage and b-precursor acute leukemias: dominance of hox dysregulation [J]. *Blood*, 2003, **102**(1): 262 - 268.
[25] Barczak A, Rodriguez M W, Hanspers K, et al. Spotted long oligonucleotide arrays for human gene expression analysis [J]. *Genome Res*, 2003, **13**(7): 1775 - 1785.

关于白血病亚型基因芯片数据资料的后综分析

杨锡南　孙　啸　陆祖宏

（东南大学生物电子学国家重点实验室，南京 210096）

摘要：针对筛选特定癌症亚型的特异表达基因，提出了一种新颖的癌症基因芯片的后综分析方法——运用并改进秩打分算法（RS），对有序基因列表的统计均值取秩并打分. 该算法结合常用的“一对多”（OVA）比对方法或“一对一”（OVO）比对方法，在跨平台检测某些白血病亚型特异基因时显得极为有效. 6 个公开的白血病数据的统计结果显示白血病亚型间的分子生物信号差异强于芯片系统间的差异. 此外，一组儿童白血病亚型（BCR_ABL）的标志基因能够准确预测成人白血病中的该亚型. 结果有助于从白血病或其他有着充分研究背景的癌症芯片表达数据中，发现、确认和治疗其亚型.

关键词：寡核苷酸生物芯片；后综分析；秩打分；白血病

中图分类号：O29；Q354