SAND2007-0699J

Gene Expression Overlap affects Karyotype Prediction in Pediatric ALL

To the Editor

Leukemia is the most common childhood malignancy in the United States (1). Acute lymphoblastic leukemia (ALL) accounts for 75% of new leukemia cases in children. Although the outcome for children with ALL has improved dramatically over the past three decades, 25% of children with ALL still develop recurrent disease. Current risk classification schemes in pediatric ALL use clinical and laboratory parameters such as age and initial white blood cell count, as well as the presence of specific ALL-associated cytogenetic or molecular genetic abnormalities. Stratification based on cytogenetic analysis and molecular genetic detection consider B precursor ALL translocations such as t(12;21)[TEL-AML1], t(1;19)[E2A-PBX1], and t(9;22)[BCR-ABL], as well as numerical imbalances such as hyperdiploidy, specific chromosome trisomies, or hypodiploidy. Despite such efforts, current diagnosis and risk classification schemes in pediatric ALL remain imprecise. In particular, it is likely that a significant number of higher-risk children are currently over-treated and could be cured with less intensive regimens, resulting in fewer toxicities and long-term side effects. Conversely, a significant number of children in lower-risk categories still relapse and precise means to prospectively identify them have remained elusive.

The advent of DNA-based microarray technology raises the possibility of improving our understanding of the pathogenesis and treatment of leukemia. There have been efforts at genome-wide studies of leukemia classification. Both Yeoh *et al.* (2) and Ross *et al.* (3) have reported that gene expression profiling can identify the known prognostic subtypes of ALL. Yeoh *et al.* have further determined that T-ALL cases can be divided into intrinsic biological clusters, with risk groupings similar to B precursor ALL. We report the results of another gene expression experiment using a differet cohort of pediatric ALL cases. Our cohort is

more typical of pediatric ALL at presentation in that it includes patients with the above well-defined karyotypes as well as patients with ambiguous cytogenetic abnormalities. Using a variety of machine learning techniques, we have validated the results of Yeoh *et al*., but have also discovered that expression profiles can overlap when considering cohorts containing a mixture of patients with and without well-defined cytogenetic abnormalities. This overlap results in difficulties predicting karyoptype from gene expression and can be seen in both supervised and unsupervised machine learning approaches, revealing complexity and novel biologic clusters not precisely correlated with the known abnormalities.

Our 311 patient cohort samples were collected with informed consent by the Pediatric Oncology Group (POG) in therapeutic trials ALinC15 and ALinC16 (8602, 9005, 9006, 9201, 9405, 9406, and 9605). The details of these clinical trials have been previously reported (4-6). The diagnosis of ALL was made on morphologic evaluation of bone marrow or peripheral blood and was confirmed by central review. Gene expression profiling was performed using RNA isolation (Qiagen RNeasy), amplification, and hybridization to HG_U95Av2 oligonucleotide microarrays (Affymetrix). The HG_U95Av2 chips were scanned and analyzed following the Affymetrix Microarray Suite (MAS) Version 5.0 Software. Quality control criterion related to percent leukemic blasts, sample quantity, RNA quality, oligonucleotide staining, array hybridization, and amplification were applied to eliminate 57 of the original 311 samples, ultimately leaving a cohort of 254. All Affymetrix microarray signal and CEL data, together with covariate clinical, cytogenetic, and annotated experimental information is available at the National Cancer Institute Cancer Array Informatics website (http://caarraydb.nci.nih.gov/caarray/: Experiment ID 1015897590271440). A detailed description of experimental protocol and quality control criterion can be found in the supplemental material.

The 254 patient dataset was first pre-processed by the removal of control probe sets (AFFX accession IDs), and probe sets with no "present" calls, as determined by the Affymetrix MAS 5.0 statistical software. After this process, 8,943 of the original 12,625 Affymetrix HG_U95Av2 probe sets were retained for analysis. Next a base-10 logarithmic transformation of the gene expression data was performed. Finally, the 254 patient dataset was divided into a 167 patient training set and an 87 patient test set. The training set was selected at random but was balanced so that the distribution of well-defined karyotypes (including t(12;21)[TEL-AML1], t(4;11)[AF4-MLL], t(1;19)[E2A-PBX1], t(9;22)[BCR-ABL], and hyperdiploid>50) was representative of the entire dataset. The final distribution of patients in the dataset, including the training/test set split is shown according to karyotype in Table 1. We note that roughly half of our dataset consists of ambiguous karyotypes, in contrast to the datasets used in both Yeoh *et al.* (2) and Ross *et al.* (3). The datasets used in Yeoh *et al*. and Ross *et al*. are dominated by well-defined karyotypic abnormalities.

We first analyzed the dataset using Principal Component Analysis (PCA) and VxInsight (7). For these methods we ignored the training/test set split and instead used all 254 patients. PCA resulted in two groups, as shown in Figure 1A. The first group (bottom left) was found to consist of T-lineage ALL, while the second group (upper right), consisted of B-lineage ALL. However, the B-ALL group was not seen to separate by karyotype, as is evident by examining the color-coded legend in Figure 1. The B-ALL group also failed to separate even excluding T-ALL, as seen in Figure 1B. Since this result did not agree with the results previously obtained by Yeoh *et al.* and Ross *et al*., we performed an additional test of our algorithms. Specifically, we formed a truncated version of the full dataset by excluding all ambiguous karyotypes. Hypodiploid was also excluded. This truncated dataset consisted of 128 patients and was re-analyzed using PCA. We found that clustering the truncated dataset did indeed reveal the karyotype clusters, as shown in Figure 1C, in agreement with the

results obtained by Yeoh *et al.* (2) and Ross *et al.* (3). The results improved further when we excluded hyperdiploid>50, as shown in Figure 1D.

We also used VxInsight to cluster the dataset. VxInsight produced nine distinct clusters. Two of these clusters accounted for the T-ALL cases previously seen in Figure 1A, while the remaining seven clusters accounted for the B-ALL cases. Analysis of the B-ALL clusters showed some trends, but no cytogenetic abnormality precisely defined any specific cluster. Cases with a t(12;21)[TEL-AML1] or hyperdiploidy, both conferring low risk and good outcomes, tended to cluster together, and the t(1;19)[E2A-PBX1] with poor outcome also clustered. However, ALL cases with t(9;22)[BCR-ABL] did not cluster and appeared to be distributed among virtually all B precursor clusters. Details on the VxInsight analysis, including visualizations (Figures S1.1-S1.2) and gene lists (Tables S1.1-S1.8) differentiating between the various clusters, can be found in the supplemental material.

We next turned to the supervised analysis of our dataset. We first identified gene lists for distinguishing the karyotypes t(12;21)[TEL-AML1], t(4;11)[AF4-MLL], t(1;19)[E2A-PBX1], t(9;22)[BCR-ABL], and hyperdiploid>50. We produced gene lists using ADA, SVM-RFE, discriminant analysis, fuzzy inference methods, and a TNoM variant used in conjunction with the Bayesian network methodology (8). The lists for each case were produced from the training set by contrasting a given karyotype against all other cases, excluding T-ALL. The lists were also used to obtain classifiers, which were validated on the test set. The performance of the classifiers is shown in Figure 2A.

We evaluated the statistical significance of the classification results using the log odds ratio. We found significance in all cases using some method and in most cases using every method (see Table S2.1 in the supplemental materials). We obtained the best results on t(12;21)[TEL-AML1] and t(1;19)[E2A-PBX1]. For the purposes of interpretation and comparison, we also combined our gene lists using a voting method. A description of the

voting method along with the combined gene lists can be found in the supplemental materials (Tables S3.1-S3.5). Although direct comparison of our combined list with the different lists found in Yeoh *et al.*(2) was difficult, we found that the percent overlap between our combined list and the lists in Yeoh *et al.*(2) was between 12-50% for t(12;21)[TEL-AML1]; 10-15% for t(4;11)[AF4-MLL]; 35-100% for t(1;19)[E2A-PBX1]; 8-33% for t(9;22)[BCR-ABL]; and 5-18% for hyperdiploid>50. More details can be found in the supplemental materials (Tables S4.1-S4.6 and Figures S2.1-S2.5).

Our supervised analysis supports our previous results using the unsupervised methods. As shown in Figure 2A, all methods achieved good accuracy when classifying a given karyotype against all other cases. However, when measured using sensitivity, defined as the ratio of true positives to true positives and false negatives (TP/(TP+FN)), we saw a decrease in performance, seen in Figure 2B. Specificity, the ratio of true negatives to true negatives and false positives (TN/(TN+FP)), was high, seen in Figure 2C. Together, these measures imply that our classifiers were accurate outside a specific karyotype, but inaccurate within that karyotype. In other words, there was a lot of overlap between karyotype predictions, so that the karyotype predictions were difficult to distinguish when seen in the context of the entire dataset. Furthermore (as we would expect), the sensitivities of the supervised methods improved when applied to the truncated dataset (data not shown).

The low sensitivity obtained by the supervised methods supports our previous results using PCA. A low sensitivity implies an overabundance of false negatives. The source of these false negatives was in our case the patients with ambiguous cytogenetics. When these patients were present, we could not differentiate the karyotypes using unsupervised learning. When the ambiguous patients and corresponding false negatives were removed, the sensitivity increased, and the unsupervised techniques were able to identify the karyotypes.

One of the primary tools for diagnosis and treatment of pediatric ALL is the assignment of patients to risk categories based on cytogenetics. The gene expression profiling work in Yeoh *et al*. (2) and Ross *et al*. (3) has been undertaken to improve the classification of the assignment of patients to those risk categories. If successful, future classification may rely more heavily on gene expression profiling than on traditional methods such as morphology, immunophenotyping, cytogenetics, and molecular diagnostics. However, we have found that gene profiling is less accurate for karyotype prediction than previously understood. Patients with ambiguous cytogenetic abnormalities may have similar gene expression profiles to patients within the major categories of well-defined abnormalities. This similarity results in a high degree of overlap when using gene expression profiles trained to recognize specific karyotypic abnormalities. This overlap may result in prediction difficulties, as shown in this study using both supervised and unsupervised machine learning approaches. Furthermore, the unsupervised approaches used in this study have revealed the existence of several novel groups of patients apparently not highly correlated with the major ALL-associated prognostically relevant karyotypes.

While our data indicates that gene expression profiling to karyotype classification may be more difficult than previously thought, this does not suggest that gene expression profiling will not be useful for diagnosis and treatment of pediatric ALL. On the contrary, our data suggest that cytogenetic-based classification schemes alone may not be sufficient for precise diagnosis and risk classification in ALL. Gene expression profiling may provide a more comprehensive classification than cytogenetics alone, as it takes into consideration the biological state of the cell and the genetic progression of the disease, perhaps eventually providing a more accurate classification for directed therapy. In this study, unsupervised learning approaches have revealed complexity and novel biologic clusters of ALL that are not precisely correlated with the frequent cytogenetic abnormalities seen in this disease. These

results imply that specific genetic lesions may be initiating events that are not sufficient for transformation. Gene expression profiling may therefore provide a more comprehensive picture of the genetic progression in this heterogeneous disease and may be more capable of grouping cases with similar functional activation or inactivation of specific biologic pathways. As we further develop insight into the functional pathways that are perturbed in each of these clusters, we may develop new insights for improved molecular classification and therapeutic targeting.

Acknowledgements

This work was supported in part by D.H.H.S. National Institute of Health Grants NCI CA88361, NCI CA32102, and NCI CA86026; grants from the W.M. Keck Foundation; the Dedicated Health Research Fund of the State of New Mexico; D.O.E. Laboratory Directed Research and Development Program Grant SAND2003-4664 at Sandia National Laboratories; and support from the UNM Cancer Center Genomics, Biostatistics, and Biocomputing Shared Facilities. Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company for the United States Department of Energy under contract DE-AC04-94AL85000.

Shawn B. Martin[1], Monica P. Mosquera-Caro[2], Jeffrey W. Potter[2], George S. Davidson[1], Erik Andries[2,3], Huining Kang[2,3], Paul Helman[3], Robert L. Veroff[3], Susan R. Atlas[3], Maurice Murphy[3], Xuefei Wang[3], Kerem Ar[2], Yuexian Xu[2], I-Ming Chen[2], Frederick A. Schultz[2], Carla S. Wilson[2], Richard Harvey[2], Ed Bedrick[3], Jonathan Shuster[4], Andrew J. Carroll[5], Bruce Camitta[6], and Cheryl L. Willman[2]

[1]Computational Biology, Sandia National Laboratories, Albuquerque, New Mexico, USA; [2]Department of Pathology and the Cancer Research and Treatment Center, University of New Mexico School of Medicine, Albuquerque, New Mexico, USA; [3]Departments of Computer Science, Mathematics and Statistics, Physics and Astronomy, and the Center for High Performance Computing, University of New Mexico, Albuquerque, New Mexico, USA; [4]General Clinical Research Center, The University of Florida, Gainesville, Florida, USA; [5]Department of Genetics, University of Alabama, Birmingham, AL, USA; [6]Department of Pediatrics, Midwest Children's Cancer Center, Medical College of Wisconsin, Milwaukee, WI, USA.

Correspondence: Dr. Cheryl L. Willman, Director and CEO, UNM Cancer Research and Treatment Center, 900 Camino de Salud N.E., Albuquerque, New Mexico 87131, USA. Phone: 1(505)272-5622, fax: 1(505)272-4039, email: cwillman@salud.unm.edu.

References

1. Pui CH, Relling MV, Downing JR. Acute lymphoblastic leukemia. N Engl J Med 2004;350(15):1535-48.
2. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 2002;1(2):133-43.
3. Ross ME, Zhou X, Song G, Shurtleff SA, Girtman K, Williams WK, et al. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood 2003;102(8):2951-9.
4. Lauer SJ, Shuster JJ, Mahoney DH, Jr., Winick N, Toledano S, Munoz L, et al. A comparison of early intensive methotrexate/mercaptopurine with early intensive alternating combination chemotherapy for high-risk B-precursor acute lymphoblastic leukemia: a Pediatric Oncology Group phase III randomized trial. Leukemia 2001;15(7):1038-45.
5. Mahoney DH, Jr., Shuster J, Nitschke R, Lauer SJ, Winick N, Steuber CP, et al. Intermediate-dose intravenous methotrexate with intravenous mercaptopurine is superior to repetitive low-dose oral methotrexate with intravenous mercaptopurine for children with lower-risk B-lineage acute lymphoblastic leukemia: a Pediatric Oncology Group phase III trial. J Clin Oncol 1998;16(1):246-54.
6. Mahoney DH, Jr., Shuster JJ, Nitschke R, Lauer S, Steuber CP, Camitta B. Intensification with intermediate-dose intravenous methotrexate is effective therapy for children with lower-risk B-precursor acute lymphoblastic leukemia: A Pediatric Oncology Group study. J Clin Oncol 2000;18(6):1285-94.

7. Davidson GS, Hendrickson B, Johnson DK, Meyers CE, Wylie BN. Knowledge mining with VxInsight: discovery through interaction. Journal of Intelligent Information Systems 1998;11(3):259-285.

8. Helman P, Veroff R, Atlas SR, Willman CL. A new bayesian network classification methodology for gene expression data. J. Comput. Biol. 2004;11(4):581-615.

Table and legend to table

Karyotype	Training Set	Test Set	Total	Percent of Dataset	Normal Population
t(12;21)[TEL-AML1]	24	6	30	11.8%	22%
t(4;11)[AF4-MLL]	14	8	22	8.7%	8%
t(1;19)[E2A-PBX1]	21	11	32	12.6%	5%
t(9;22)[BCR-ABL]	10	5	15	5.9%	3%
Hyperdiploid>50	17	12	29	11.4%	25%
Hypodiploid	2	2	4	1.6%	1%
Ambiguous	79	43	122	48.0%	36%
Total	167	87	254	100.0%	100%

Table 1. Distribution of ALL patients according to karyoptype. The Percent of Dataset column shows the percent (including both training and test set) of a given karyotype in the dataset, while the Normal Population column shows the percent of that karyotype expected in a clinical setting (1).

Titles and legends to figures

Figure 1. Prinicpal Component Analysis. The results of PCA vary with the number of patients used. On the upper left (A), a PCA clustering using all patients puts T-ALL on the lower left and B-ALL on the upper right (multiple colors according to karyotype). On the upper right (B), PCA fails to resolve the B-ALL patients even excluding T-ALL. In the lower left (C), PCA resolves the B-ALL cluster according to karyotype when ambiguous karyotypes are excluded from the dataset. On the lower right (D), we included only the translocations.

Figure 2. Performance of karotype classifiers. From left to right (A-C) these plots show the accuracy, sensitivity, and specificity of the various classifiers trained to distinguish karyotypes. The key on the far right indicates the classification methods used.

Figures

Figure 1.

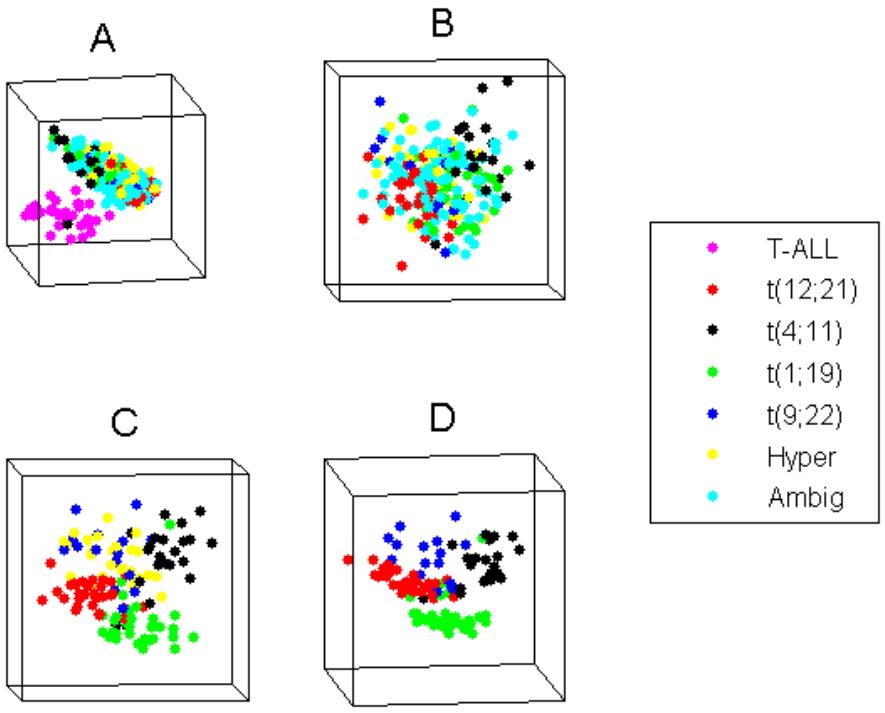

Figure 2.