

# Peeling Off the Hidden Genetic Heterogeneities of Cancers Based on Disease-Relevant Functional Modules

Jian-zhen Xu,<sup>1\*</sup> Zheng Guo,<sup>1,2\*</sup> Min Zhang,<sup>1</sup> Xia Li,<sup>1,2</sup> Yong-jin Li,<sup>1</sup> Shao-qi Rao<sup>1,3</sup>

<sup>1</sup>Department of Bioinformatics, Harbin Medical University, Harbin, China; <sup>2</sup>School of Biological Science and Technology, Tongji University, Shanghai, China; <sup>3</sup>Department of Molecular Cardiology and Department of Cardiovascular Medicine, the Cleveland Clinic Foundation, Cleveland, OH, USA.

Discovering molecular heterogeneities in phenotypically defined disease is of critical importance both for understanding pathogenic mechanisms of complex diseases and for finding efficient treatments. Recently, it has been recognized that cellular phenotypes are determined by the concerted actions of many functionally related genes in modular fashions. The underlying modular mechanisms should help the understanding of hidden genetic heterogeneities of complex diseases. We defined a putative disease module to be the functional gene groups in terms of both biological process and cellular localization, which are significantly enriched with genes highly variably expressed across the disease samples. As a validation, we used two large cancer datasets to evaluate the ability of the modules for correctly partitioning samples. Then, we sought the subtypes of complex diffuse large B-cell lymphoma (DLBCL) using a public dataset. Finally, the clinical significance of the identified subtypes was verified by survival analysis. In two validation datasets, we achieved highly accurate partitions that best fit the clinical cancer phenotypes. Then, for the notoriously heterogeneous DLBCL, we demonstrated that two partitioned subtypes using an identified module ("cellular response to stress") had very different 5-year overall rates (65% vs. 14%) and were highly significantly ( $P < 0.007$ ) correlated with the clinical survival rate. Finally, we built a multivariate Cox proportional-hazard prediction model that included 4 genes as risk predictors for survival over DLBCL. The proposed modular approach is a promising computational strategy for peeling off genetic heterogeneities and understanding the modular mechanisms of human diseases such as cancers.

Online address: <http://molmed.org>

doi: 10.2119/2005-00036.Xu

## INTRODUCTION

Genetic heterogeneity describes the biological complexities whereby apparently similar inheritable characters result from different genes or different genetic mechanisms. In clinical settings, genetic heterogeneity refers to the presence of a variety of genetic defects that cause the same disease as defined in the current disease classifications (1), a finding common to a list of complex human diseases such as cardiovascular disease, cancer, diabetes, autoimmunity, psychiatric ill-

ness, and many others, and even Mendelian disorders (2). Genetic heterogeneity has profound influences on modern clinical practice and biomedical research of common human disease. In the basic genomic sciences, it is a thorny issue for genetic linkage analysis (3,4), high-density admixture mapping of disease genes (5), and microarray data analysis (6). More accurate phenotyping of genetic heterogeneous samples, either by explicitly modeling stratified population structure [for example, due to racial difference (7)] or by peeling off hidden ge-

netic heterogeneity (4), has been demonstrated to result in increased power to map disease genes. In clinical practice, it is increasingly recognized that our current categorization of human diseases still lumps together molecularly distinct diseases (for example, cancers) with the same clinical phenotypes (8). Because the clinical behaviors of some complex diseases such as cancers cannot be accounted for completely by morphological or pretreatment clinical characteristics, patients with the same phenotype, which might be caused by different underlying molecular mechanisms, often show different responses to drug treatment and have different prognoses. Thus, a central challenge to study and to improve efficacy in treatment of complex diseases is to resolve their molecular heterogeneity mechanisms (9).

Genomic-scale molecular data rapidly accumulating from biomedicine domains offer opportunities to peel off genetic

---

**Address correspondence and reprint requests to Zheng Guo, Department of Bioinformatics, Harbin Medical University, Harbin, China 150086. Phone: +86-451-8661-5933; fax: +86-451-8666-9617; e-mail: guoz@ems.hrbmu.edu.cn; or Shaoqi Rao, Department of Molecular Cardiology/Office NB5-28, Lerner Research Institute /NB50, Cleveland Clinic Foundation, 9500 Euclid Ave., Cleveland, OH 44195. Phone: 216-444-0056; fax: 216-444-2682; e-mail: raos@ccf.org.**

\*J.X. and Z.G. contributed equally to this study.

Submitted November 11, 2005; accepted for publication March 22, 2006.

heterogeneities at the molecular level. The promise for molecular classification and discovery of hidden disease subtypes has been realized in successful stratification of diffuse large B-cell lymphoma (DLBCL) (8), based on the expression profiles of thousands of genes measured by microarrays and using computationally clustering algorithms, an approach aimed at defining genetically homogeneous novel cancer subtypes among cancer patients. Although traditional clustering analysis of the expression profiles of individual genes is a successful approach to discovering disease subtypes, several significant shortcomings in this analysis strategy remain. First, a traditional clustering analysis often groups patients with overall similar gene expression profiles by the complete set of thousands of genes represented on the arrays, and thus has low ability to reflect the influence of the most disease-relevant genes. When a large number of irrelevant or weakly relevant genes greatly influence the clustering results, spurious structure of disease expression patterns may appear out of the high dimensionality of data. Second, and more important, because a traditional clustering analysis rarely uses the current gene functional knowledge to the groupings, the biological relevance and interpretation of the patient groupings by the traditional clustering analysis are often unclear. It is obviously of great value if we can discover and elaborate disease subtypes at the functional module level by explicitly yielding functionally compact gene sets with coherent expression across cancer samples.

A module describes a biologically coherent set of genes that tend to express and perform their highly integrated cellular functions in somewhat isolated and interactive modular fashions (10,11), a phenomenon that has inspired studies for elucidating the high-order pathogenic mechanisms of complex diseases (12,13). For example, Mootha et al. (12) showed that the modest but coordinate disease-associated changes of a set of functionally related genes could be identified

even in the cases where the expression of individual genes was not significantly different. Segal et al. (14) defined “modules” as biologically meaningful gene sets that are conditionally activated or repressed across a wide variety of cancer types, and identified some modules deregulated in cancer. Our recent study demonstrated that cancer types can be precisely and robustly classified based on functional modules enriched with differentially expressed genes (15). Nevertheless, nothing in the literature exists for fully exploiting the power and value of the modular approaches to systematically dissecting the molecular heterogeneities of human diseases.

Here, we further proposed a module-based clustering approach for dissection of cancer heterogeneity by using the disease-relevant functional modules. First, we selected differentially expressed genes under the disease conditions. It should be noted that algorithms such as *t* test or *F* test are not proper for selecting genes under the disease heterogeneity (subtypes), because the validity of these tests relies on accuracy in describing the disease population structure by the current clinical disease categorizations, i.e., lack of phenotypic heterogeneity. Hence, we took a robust metric, the overall variability of gene expressions, to guide gene selection. Genes with top-ranked expression variations across samples, which explain most of the total variances potentially contributed by known or unknown factors (for example, the hidden disease subtypes), were selected as “feature genes.” This metric has been adapted by several researchers for initial gene selection (16,17). Then, we identified cellular-localized biological processes enriched with feature genes as “putative signature modules.” Finally, we partitioned samples to seek for hidden disease subtypes using the expression profiles of the genes annotated to these well-characterized modules. As subcellular localization of genes and proteins is a key functional characteristic determining their ability to interact with other proteins and small metabolites in their local environment,

we characterized the modules in terms of biological processes and cellular localizations based on Gene Ontology (GO) (18). GO is a comprehensive ontological system describing gene functions in three directed acyclic graphs: biological process (BP), molecular function (MF), and cellular component (CC). In numerical analyses, we first validated the proposed modular approach for accurately partitioning cancer phenotypes using two publicly available large cancer datasets. Then, we used the approach to explore the hidden subtypes of a notoriously heterogeneous phenotype, DLBCL (8). The results demonstrated that two partitioned subtypes using an identified functional module had very different 5-year overall rates, and the partition was highly significantly correlated with the clinical survival rate.

## MATERIALS AND METHODS

### Description of Datasets

We used two large datasets to evaluate the goodness-of-fit performance of the proposed modular approach. The liver cancer dataset (19) consists of 23,075 cDNAs measured in 105 primary hepatocellular carcinoma (HCC) samples and 76 normal liver tissues, a typical large disease-control example. Because the HCC phenotype is specifically defined, it can be reasonably assumed that the original tissue phenotypes were well characterized. We further explored the reliability of the proposed modular approach for partitioning between various types of cancers by analyzing a classical multiple-class dataset, NCI60 (20), which consists of 9,703 cDNAs measured in 60 cell lines of 9 cancer types. The data for non-small cell lung carcinoma and breast tumors were not used in this study because of the possible existence of heterogeneous hidden subtypes (20) or misassigned labels (21) for their samples. The data for prostate cancer were also excluded because they consisted of only 2 samples. Thus a subset of the NCI60 data (41 samples of 6 cancer types) was used in this study, including 8 samples of renal can-

cer (RE), 7 of colon cancer (CO), 6 of leukemia (LE), 8 of melanoma (ME), 6 of ovarian cancer (OV), and 6 of central nervous system cancer (CNS). After evaluating its ability for accurately partitioning the two diverse data structures, we used the proposed modular approach to peel off the hidden subtypes of DLBCL, which has been demonstrated to be notoriously heterogeneous (20,22,23). The third dataset consists of 4,026 cDNAs measured in 42 DLBCL samples (8). We verified the identified hidden partitions (DLBCL subtypes) by survival analysis of the clinical profiles of patients in each molecular-based partition.

For each of the above cDNA microarray datasets, we screened out clones with missing data in more than 5% of arrays and applied a base-2 logarithmic transformation. As in Alizadeh et al. (8), we imputed remaining missing data with zeros. Each experiment was standardized to zero median across the genes. The datasets of HCC, NCI60, and DLBCL finally comprised 10,516, 6,748, and 2,751 genes, respectively.

### Selecting Putative Signature Modules from Gene Ontology

Most current approaches to defining modules use only BP categorization of GO. However, a BP category may actually encompass the genes involved in distinct processes occurring in different cellular compartments, and genes even within the same BP may show a clear expression distinction with respect to their subcellular localizations (24,25). Therefore, to identify modules containing the consistently coexpressed genes potentially aroused by the disease conditions, we sorted genes of a BP category into CCs to form combined categories. For example, genes whose protein products function in cell adhesion (BP) on membrane (CC) are accommodated in a combined GO category. We referred to all the measured genes annotated to at least one of the combined categories as “annotated genes.”

For each dataset, the top  $x$  percent of genes with the largest expression vari-

ances were selected as feature genes. Then, we used a hypergeometric distribution (26,27) to calculate the probability  $P$  of a combined GO category having the number of annotated feature genes by chance; a smaller  $P$  value corresponds to a higher likelihood of the feature genes enriched in the category. We selected categories with  $P \leq 0.001$  and kept the categories containing at least 30 feature genes to retain enough data for clustering. Owing to the hierarchical nature of the GO-structured categories, there are some redundancies in the selected categories; for example, BP function description is the same but a general-specific (for example, parent-child) relationship lies on the CC functional description. In such a case, only the combined category with the child category in cellular component ontology was reserved, because its functional description is more specifically defined. In the following text, we refer to such GO categories as a “module” for short. The identified modules should be statistically robust to the differences in the criteria for selecting feature genes because the analysis results are determined by the joint statistical behaviors of sets of genes (27). We demonstrated the robustness of the modules by comparing the modules identified at different top percentage levels ( $x = 10, 15, 20$ ) of feature genes with the largest variances.

### Clustering Samples Based on Individual Modules

For each identified module, we extracted the expression profiles of the measured genes that were annotated to it. By agglomerative hierarchical clustering (28), each sample was initially assigned to one cluster, then the distances between all clusters were computed and the two clusters with the smallest distance value were merged. Distance computation and merging were repeated until there was only one cluster left. In this work, Pearson correlation was used for the distance metric, and the centered average linkage method was used for merging. For the purpose of evaluating

the modular clustering approach, we adopted the predefined cluster number in the original data source for pruning off the hierarchical tree and allocating the samples into clusters. Because the expected value of the Rand index is not constant for random partitions (29), we used the adjusted Rand index (ARI) (30) to measure the agreement between the identified clusters and the original partitions (for example, the clinical sample labels). The expected value of the ARI is 0 when the partitions are drawn randomly, and the ARI is 1 when two partitions agree perfectly. A larger ARI dictates a higher correspondence between two types of partitions.

One general approach to assess the significance of the observed ARI for a module might be to compare the ARI value with those of the same-sized gene subsets randomly selected from the whole microarrays that contained the genes (and their coexpressed ones) in the current modules. However, we are more interested in finding whether the profiles of the genes in the current modules were significantly better at clustering than the gene groups randomly selected from a null (or contrast) population, where the gene had no or less functional relationship with the current modules. It is well known that similarly expressed (coexpressed) genes tend to share the same or similar functions (31,32), and in fact the gene coexpression information is often used for predicting gene functions (33,34). Thus, we constructed the null gene population using the silence genes among all the annotated genes from the original expression profiles, after excluding (i) the genes annotated to the identified modules and (ii) the genes significantly coexpressed with at least one gene in the identified modules. Here, two genes were defined as coexpressed when the absolute value of Pearson correlation coefficient ( $\gamma$ ) of their expressions was larger than a threshold corresponding to the significance level  $P \leq 0.005$ , determined by using 10,000 gene pairs randomly sampled from the original expression profiles.

Then, for each identified module, 1000 gene subsets of the same size as the module were randomly sampled from the null population. Applying the same clustering procedure to the 1000 random gene subsets, we set the  $P$  value of the ARI of the module as the fraction of 1000 random subsets having ARIs larger than that of the module. The  $P$  value based on such randomizations was used to assess whether the observed ARI for a module was achieved by chance or, in a more specific sense, whether the module was better at clustering (that is, more likely relevant to the phenotypic partitions) than gene subsets that were less likely to be of close functional relationship with the identified modules.

### Clustering Based on Multiple Modules

Some samples can be possibly misallocated by using one or a few modules. One robust way to get improved partition results is to decide samples' labels in a collectively voting manner, by fusing the results from the individual modules. Here, for each sample, based on its membership labels obtained from different modules, we applied a simple majority rule to determine a sample's membership. If the sample had the highest votes across several classes, we randomly assigned one of the class labels to the sample.

### Survival Analysis

To verify the clinical significance of the identified hidden DLBCL subtypes, we estimated survival curves by Kaplan-Meier product-limit method and assessed the differences between the survival curves of the subtypes of DLBCL patients by log-rank test (35). To construct a model for predicting the overall survival time, univariate Cox proportional-hazards model (36) was used to determine the significance (at  $P \leq 0.05$ ) of the effects of the genes annotated to the identified module or modules on the patients' survival months. Subsequently, genes past the above threshold were re-analyzed using a multivariate Cox proportional-hazards regression model, with

the overall survival months as the dependent variable. Wald  $\chi^2$  test was used to determine the significance of each predictor's hazard toward the survival time.

## RESULTS

### Validation of the Proposed Modular Approach Using Two Large Microarray Datasets

In the liver cancer dataset, we identified 41 combined categories significantly ( $P \leq 0.001$ ) enriched with 10% top-ranked genes with the largest expression variances. When two combined categories had the same BP function description and their CC descriptions were of a general-specific relationship, only the combined category with the more specific CC description was retained. For example, the combined category, "BP: development" and "CC: cellular component," was removed because a more specifically defined module ("BP: development" and "CC: extracellular") could be identified. It should be noted that the purpose for removing some redundant modules was the promise of finding a set of more compact and more specific functional modules that would provide sufficient information for characterizing cancer samples. The redundant modules that overlapped in one or two dimensions with the selected ones were often highly enriched with feature genes, too, suggesting that they were also good candidates for separating disease samples. For example, based on the gene expression profiles in the module of "BP: development" and "CC: cellular component," an ARI of 0.732 was obtained.

After the redundancy treatment, six modules were left for the following analysis. We used the expression profiles of the measured genes annotated in each of the six modules to partition the samples. The clustering results based on each of the six modules agreed well with the original clinical labels, and the observed ARIs were 0.830, 0.871, 0.892, 0.790, 0.713, and 0.850. The average ARI for the six modules was 0.824 ( $\pm 0.065$ ), and the module "BP: cell growth and/or

maintenance" occurring at "CC: extracellular" achieved the best results, with ARI 0.892. Then, for each module, we randomly selected 1000 gene subsets of the same size of the module from the null population as described previously. We found that no random subset achieved an ARI larger than that of the corresponding module, so the observed ARIs of all six modules were significantly ( $P < 0.001$ ) better at clustering than randomly selected gene subsets. The sample memberships assigned by the six individual modules show that some samples were misallocated by one or more modules (Supplement 1). By using majority rule clustering, which assigns the majority membership labels to samples, we obtained an ARI of 0.934, where only 3 tumor samples were misallocated (Table 1).

Accumulated biological experiments provided rich evidence to support the roles of some key proteins annotated to the six modules. For example, it has been reported that nucleoside transporters and glutamine transporters are abnormally expressed in hepatoma cells (37,38), which supports that the module of "BP: transport" occurring at "CC: integral to plasma membrane" is relevant to HCC. The significant correlations of serum IL-8 levels with tumor size and tumor stage (39) suggest that two modules ("BP: G-protein coupled receptor protein signaling pathway" and "BP: immune response," both occurring at "CC: extracellular region") may be directly or indirectly involved in the progression of HCC. Genes such as vascular endothelial growth factor (*VEGF*), annotated to the module "BP: cell development" and "BP: signal pathway" occurring at "CC: extracellular region," have been suggested as diagnostic markers or prognostic factors of HCC (40). In addition, glypican 3 (*GPC3*) (in module "BP: G-protein coupled receptor protein signaling pathway" occurring at "CC: cell") has been found to be both a marker for HCC and a target for HCC therapy (41).

Based on the NCI60 dataset, we identified 38 combined categories signifi-

**Table 1.** Selected modules for liver cancer dataset and NCI60 dataset.

Datasets	BP category <sup>a</sup>	CC category <sup>b</sup>	<i>P</i> (module) <sup>c</sup>	<i>N</i> <sup>d</sup>	ARI <sup>e</sup>	<i>S</i> <sup>f</sup>	<i>P</i> (ARI) <sup>g</sup>
Liver cancer	GO:0006955: immune response	GO:0005576: extracellular	< 2.22E-16	82	0.830	8	< 0.001
	GO:0007275: development	GO:0005576: extracellular	< 2.22E-16	127	0.871	6	< 0.001
	GO:0008151: cell growth and/or maintenance	GO:0005576: extracellular	2.90E-11	139	0.892	5	< 0.001
	GO:0007165: signal transduction	GO:0005576: extracellular	7.78E-10	104	0.790	10	< 0.001
	GO:0006810: transport	GO:0005887: integral to plasma membrane	1.05E-06	168	0.713	14	< 0.001
	GO:0007186: G-protein coupled receptor protein signaling pathway	GO:0005623: cell	1.05E-04	133	0.850	7	< 0.001
Majority rule	NA <sup>h</sup>	NA	NA	NA	0.934	3	< 0.001
Top150_liver	NA	NA	NA	150	0.871	6	< 0.001
NCI60	GO:0007155: cell adhesion	GO:0005886: plasma membrane	1.07E-08	77	0.585	8	< 0.001
	GO:0009653: morphogenesis	GO:0016020: membrane	6.30E-08	112	0.481	14	0.010
	GO:0007155: cell adhesion	GO:0016021: integral to membrane	1.01E-07	87	0.649	11	< 0.001
	GO:0007275: development	GO:0016021: integral to membrane	2.25E-06	104	0.596	14	< 0.001
	GO:0007165: signal transduction	GO:0016021: integral to membrane	6.42E-05	180	0.715	10	< 0.001
	GO:0007166: cell surface receptor linked signal transduction	GO:0016021: integral to membrane	2.13E-04	106	0.580	11	< 0.001
	GO:0006955: immune response	GO:0005623: cell	8.16E-04	133	0.640	9	< 0.001
Majority rule	NA	NA	NA	NA	0.697	8	< 0.001
Top150_NCI	NA	NA	NA	150	0.728	7	< 0.001

<sup>a</sup>The biological process description of the module; <sup>b</sup>the cellular component description of the module; <sup>c</sup>statistical significance of the selected module; <sup>d</sup>number of annotated genes; <sup>e</sup>adjusted Rand index; <sup>f</sup>number of misallocated samples; <sup>g</sup>statistical significance of ARI with exclusion of genes; <sup>h</sup>not available.

cantly ( $P \leq 0.001$ ) enriched with the 10% top-ranked genes with the largest expression variances. After the redundancy treatment, seven modules remained. All the ARIs of the seven modules were significantly larger than those achieved by chance ( $P < 0.001$  for six modules and  $P < 0.010$  for one), and the average ARI of the seven modules was 0.607 ( $\pm 0.073$ ). The majority rule clustering approach achieved an ARI value of 0.697. Detailed results are listed in Table 1 and Supplement 2. Numerous reports (42,43) have documented the relationships between cancers and the seven selected modules: cell communication (modules “signal transduction” and “cell-surface receptor linked signal transduction” and “cell adhesion”), im-

mune response, cell development (modules “development” and “morphogenesis”), and so on.

In each dataset, based on the feature genes selected as the top 10%, 15%, and 20% ranked genes with the largest variances, the identified modules largely overlapped, suggesting the robustness of such modules to the differences of the thresholds for selecting feature genes. In fact, for liver cancer, compared with the results found when  $x = 10$ , two additional modules (“defense response” and “cell-surface receptor linked signal transduction”) were identified when  $x = 15$ , and only one more module (“response to wounding”) was identified when  $x = 20$ . Similar trends were found in the NCI60 dataset.

Among the top 150 genes (about the average size of the modules across this study) with the largest variances in the liver cancer or NCI60 datasets, there were 120 and 130 genes, respectively, co-expressed with at least 1 gene in the identified modules at the significance level  $P \leq 0.005$ , determined using 10,000 gene pairs randomly sampled from the original expression profiles. Thus, we expect that the set of the top-ranked genes could achieve good clustering results. In fact, the ARIs for the set of the top-ranked 150 genes were estimated to be 0.871 and 0.728 for the two datasets (liver cancer and NCI60), respectively, which were comparable to those obtained using the majority rule modular approach (Table 1).

**Peeling Off the Hidden Genetic Heterogeneities of DLBCL**

Molecular heterogeneity in DLBCL patients was extensively investigated previously [for example, (8,22,23)]. Inspired by the successful results for partitioning HCC and NCI60 datasets, we then applied the proposed modular approach to uncover the underlying molecular subtypes of DLBCL. Based on the DLBCL dataset, six modules were identified, as shown in Table 2 and Supplement 3. The most significant module (annotated with 173 genes, and  $P \leq 9.25E-07$ ) was “GO:0006950: cellular response to stress” occurring at “GO:0005623: cell.” By this module, two distinct DLBCL subtypes were discovered via unsupervised clustering of the DLBCL patients based on the expression profiles of the annotated genes. To elucidate the clinical implications of the identified molecular module, we studied survival profiles for the two subtypes (Figure 1A). As the data of the survival months were not available for 2 patients in the original dataset, our results were based on the remaining 40 patient samples. For comparison, we also gave the Kaplan-Meier curves of overall survival for the phenotypic partitions revealed in the original clinic labels (Figure 1B). The results demonstrated that the partitions identified by the “response to stress” module had very different 5-year overall rates (65% vs. 14%), and these partitions were highly significantly ( $P = 0.007$ ) correlated with the clinical survival rate. The original partitions (clinic labels), though also highly significantly ( $P = 0.010$ ) correlated with the survival data, had a markedly lower caliber to map their differential survival profiles. It was also noted that both the partitions defined by the majority rule of the six identified modules and those defined by the counterfactual module of 150 top-ranked genes were significantly correlated with the clinical survival time, but with less significant values ( $P = 0.019$  and  $0.012$ , respectively). The results imply that these modules (or groups of genes)

**Table 2.** Significant modules for DLBCL.

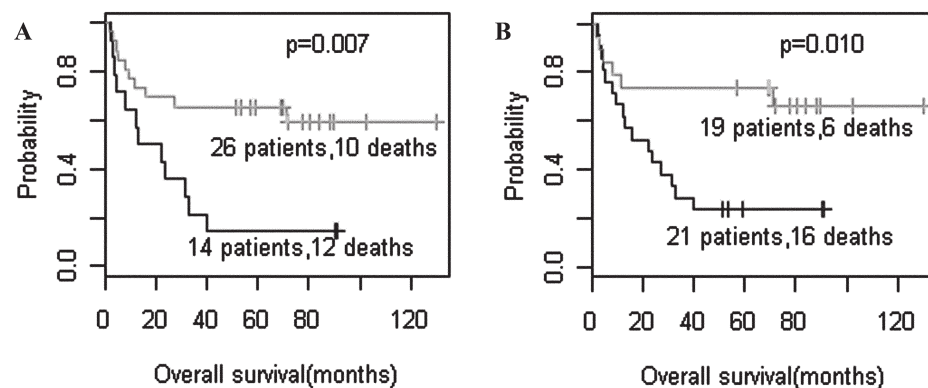
BP category <sup>a</sup>	CC category <sup>b</sup>	P (module) <sup>c</sup>	N <sup>d</sup>
GO:0006950: response to stress	GO:0005623: cell	9.25E-07	173
GO:0006955: immune response	GO:0005623: cell	1.66E-06	145
GO:0006952: defense response	GO:0005623: cell	3.81E-06	160
GO:0050874: organismal physiological process	GO:0005623: cell	6.35E-06	174
GO:0009607: response to biotic stimulus	GO:0005623: cell	3.40E-05	177
GO:0050896: response to stimulus	GO:0005623: cell	1.39E-04	279

<sup>a</sup>The biological process description of the module; <sup>b</sup>the cellular component description of the module; <sup>c</sup>statistical significance of the selected module; <sup>d</sup>number of measured annotated genes.

might have revealed the molecular heterogeneity of DLBCL from different aspects or different pathways.

To explore a compact model for clinical use, we selected a gene subset of high prediction power. Multivariate Cox proportional-hazards model was used to analyze the genes in the module labeled “response to stress.” To reduce the number of variables to be modeled, first, we applied a univariate Cox proportional-hazards model to identify the genes whose marginal effects on the overall survival time were significant. Fourteen genes (*BCL2*, *CHES1*, *ERCC5*, *HMGB2*, *IRF4*, *LY64*, *SMAD7*, *OGG1*, *RPA3*, *TNF*, *CD83*, *PDIR*, *TLK2*, and *FLJ10858*) were found at the significance level of 0.05. Then, using the stepwise variable selection option (with the same inclusion and

exclusion  $P$  value of 0.05) for the multivariate Cox proportional-hazards regression model (36), we ended up with 4 predictors (genes) (Table 3). Two of the 4 genes, cell CLL/lymphoma 2 (*BCL2*) and tumor necrosis factor (*TNF*), were previously reported as the prognostic factors for lymphoma (44,45). It is interesting to note that high-mobility group box 2 (*HMGB2*), a member of the nonhistone chromosomal high-mobility group protein family, conferred a high hazard ratio (20.04, with 95% CI 3.53-113.88) (see Table 3). A previous study (46) demonstrated that *HMGB2* has the potential to control cell- and promoter-specific down- or upregulation of in vivo transcriptional activity of different members of the tumor suppressor gene *p53* family. Another predictor gene encoding CD83



**Figure 1.** Clinically distinct DLBCL subtypes defined by gene expression profiling. (A) Kaplan-Meier plot of the overall survival of DLBCL patients grouped on the basis of gene expression profiling in “cellular response to stress” module. (B) Kaplan-Meier plot of the overall survival of DLBCL patients grouped from original clinic labels.

antigen was also found at elevated levels in 20% of chronic lymphocytic leukemia (CLL) and 5 of 7 mantle cell lymphoma (MCL) patients (47), suggesting its functional and/or prognostic significance in hematologic malignancies, particularly CLL and MCL.

Please note that supplementary information is available on the Molecular Medicine website ([www.molmed.org](http://www.molmed.org)).

## DISCUSSION

In this article, we proposed a modular-based clustering approach to find disease subtypes based on modules defined by cellular-localized biological processes. As evaluated by the liver cancer and NCI60 datasets, based on a few measured genes in an individual module, the sample partitions agreed well with the original clinical labels. We thus deem that the disease-relevant module may depict one of the multiple functional facets leading to the molecular pathogenic mechanisms. Further studying the functional descriptions of the identified modules suggests that these modules enjoy explicit relevancy to the current understanding of disease mechanisms and thus are appealing for dissecting the underlying genetic heterogeneity of cancers at the modular level. It should be noted that the proposed approach is also an efficient unsupervised feature selection method that yields multiple feature gene sets (i.e., genes annotated to the modules) of functional compactness. The genes with top-ranked expression variations across samples are selected as the initial feature genes (16,17), and then are further filtered or organized by functional modules. In general, because the selected feature genes by modular approach contain both the gene expression signatures and the functional module signatures of disease subtypes, they may provide functional guidance in experimental investigation of the pathogenesis of the studied diseases.

It has been shown that using multiple 2-dimensional characterized modules individually or jointly could achieve comparable excellent partitioning results, in-

**Table 3.** Multivariate Cox proportional-hazard analysis based on signature genes relevant to survival time.

Variable	Estimated coefficient	Wald $\chi^2$	P value	Hazard ratio (95% CI)
<i>BCL2</i>	-1.25	8.87	0.0029	0.29 (0.13-0.65)
<i>HMGB2</i>	3.00	11.43	0.0007	20.04 (3.53-113.88)
<i>TNF</i>	1.68	8.50	0.0035	5.36 (1.73-16.57)
<i>CD83</i>	2.04	6.82	0.0090	7.70 (1.67-35.60)

dicating that multiple molecular pathways may be involved in the complex disease mechanisms. In addition, our previous study (15) for classifying cancers using 1-dimensional (BP) characterization of modules demonstrated that the modular approach to using the derived modular functional expression profiles is a powerful and robust alternative approach to analyzing high-dimensional gene profiles of cancers. Although both 1- and 2-dimensional modular categorization can perform equally well, we recommend using 2-dimensional (BP and CC) characterization of modules to achieve more compact and detailed knowledge in both functionality and cellular location, data that are more useful and revealing for further experimental investigation (for example, by molecular trafficking techniques).

In supervised classification, the choice of the best module or modules for disease prediction should be relatively easy; because the sample labels in training set are given, the high accuracy rates of the classifiers trained on the modules might be used to filter more specific and critical modules highly relevant to disease pathogenesis. In unsupervised clustering analysis, however, the ARI for evaluating a clustering algorithm cannot be applied directly to choose the best module, because no cross-validation can be done. Nevertheless, according to the results in this study, some general guidelines can be given for choosing one or more modules for clustering analysis of diseases. One way is to focus on one or more biologically highly related modules to explore a specific functional facet that may correspond to a unique genetic pathway. Although this simple strategy may not

get the highest ARI, it has the advantage of focusing on specific disease mechanisms. Alternatively, disease samples may be best partitioned based on "collectively voting" from the identified modules, but with the loss of the detailed functional characterization that each module provides. Another way to use the information from the identified modules in a collective manner is to put together all the measured genes contained in the modules for clustering analysis. For example, the ARI values achieved by this approach for the liver cancer and NCI60 datasets were 0.871 and 0.622, respectively (Supplement 4).

Some early studies attempted to find cancer subtypes based on expression profiles of the genes grouped by a clustering algorithm (8). The underlying assumption is that genes with similar expression patterns are more likely to have similar biological functions, but a clustering algorithm itself does not provide proof of the best grouping of genes in terms of biological functions (48). Thus, the biological interpretation of the disease clustering results relies heavily on expert knowledge, which is often subjective (49). Here we directly used an external annotation database such as Gene Ontology to extract multiple functionally compact and coherent gene sets (modules). The application of the proposed modular approach to peel off DLBCL identified two hidden subtypes. In terms of the well-characterized modular functionality and based on the significant different survival results for the patients defined by the two hidden subtypes, the proposed computational approach is a feasible and promising toolbox for peeling off mo-

lecular heterogeneities of complex human diseases.

In this study, we took the known cluster number suggested by preassigned labels as the basis to assess the validity of the proposed approach. Although the clustering results provided good fits to the known phenotypic partitions, the assumption of the lack of heterogeneity in the two studied datasets might not be true. Likewise, the problem to estimate the correct number of clusters for peeling off hidden disease subtypes is largely unsolved. Recently, some methods for obtaining the best number of sample partitions by optimizing some validity indices have been published (50,51), which would provide additional insights on improving the proposed modular approach. By its nature, an extension of the proposed modular approach could also further refine the functional modules by integrating multiple sources of functional information at different molecular levels.

#### ACKNOWLEDGMENTS

This work was supported in part by the National High Tech Development Project of China (grant nos. 2003AA2Z2051 and 2002AA2Z2052), the National Natural Science Foundation of China (grant nos. 30170515, 30370388, 30370798, 30570424, and 30571034), the 211 Project, the Tenth "Five-year" Plan, Harbin Medical University, and the Heilongjiang Province Department of Education Outstanding Overseas Scientist grant (grant no. 1055HG009).

#### REFERENCES

1. Rieger R, Michaelis A, Green MM. (1991) *Glossary of Genetics: Classical and Molecular*. Springer Verlag, Berlin, New York.
2. Krakow D et al. (2004) Mutations in the gene encoding filamin B disrupt vertebral segmentation, joint formation and skeletogenesis. *Nat. Genet.* 36:405-10.
3. Altmuller J et al. (2005) Phenotypic and genetic heterogeneity in a genome-wide linkage study of asthma families. *BMC Pulm. Med.* 5:1.
4. Shannon WD, Province MA, Rao DC. (2001) Tree-based recursive partitioning methods for subdividing sibpairs into relatively more homogeneous subgroups. *Genet. Epidemiol.* 20:293-306.
5. Hoggart CJ, Shriver MD, Kittles RA, Clayton DG,

- McKeigue PM. (2004) Design and analysis of admixture mapping studies. *Am. J. Hum. Genet.* 74:965-78.
6. Li X, Rao S, Wang Y, Gong B. (2004) Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling. *Nucleic Acids Res.* 32:2685-94.
7. Patterson N et al. (2004) Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* 74:979-1000.
8. Alizadeh AA et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403:503-11.
9. Golub TR et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531-7.
10. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. (1999) From molecular to modular cell biology. *Nature* 402:C47-52.
11. Segal E, Friedman N, Kaminski N, Regev A, Koller D. (2005) From signatures to models: understanding cancer using microarrays. *Nat. Genet.* 37 Suppl:S38-45.
12. Mootha VK et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34:267-73.
13. Huang E et al. (2003) Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat. Genet.* 34:226-30.
14. Segal E, Friedman N, Koller D, Regev A. (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet.* 36:1090-8.
15. Guo Z et al. (2005) Toward precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics* 6:58 doi:10.1186/1471-2105-6-58.
16. Ding CH. (2003) Unsupervised feature selection via two-way ordering in gene expression analysis. *Bioinformatics* 19:1259-66.
17. Dudoit S, Fridlyand J. (2003) Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* 19:1090-9.
18. Harris MA et al. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32:D258-61.
19. Chen X et al. (2002) Gene expression patterns in human liver cancers. *Mol. Biol. Cell* 13:1929-39.
20. Ross DT et al. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* 24:227-35.
21. Ellison G, Klinowska T, Westwood RF, Docter E, French T, Fox JC. (2002) Further evidence to support the melanocytic origin of MDA-MB-435. *Mol. Pathol.* 55:294-9.
22. Monti S et al. (2005) Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood* 105:1851-61.
23. Rosenwald A et al. (2002) The use of molecular

- profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.* 346:1937-47.
24. Jimenez JL, Mitchell MP, Sgouros JG. (2003) Microarray analysis of orthologous genes: conservation of the translational machinery across species at the sequence and expression level. *Genome Biol.* 4:R4.
25. Zhou X, Kao MC, Wong WH. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci. U. S. A.* 99:12783-8.
26. Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA. (2003) Global functional profiling of gene expression. *Genomics* 81:98-104.
27. Hosack DA, Dennis G Jr, Sherman BT, Lane HC, Lempicki RA. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.* 4:R70.
28. Jain A, Dubes R. (1988) *Algorithms for Clustering Data*. Prentice Hall, New York.
29. Milligan GW, Cooper MC. (1986) A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behav. Res.* 21:441-58.
30. Hubert L, Arabie P. (1985) Comparing partitions. *J. Classification* 2:193-218.
31. Pavlidis P, Lewis DP, Noble WS. (2002) Exploring gene expression data with class scores. *Pac. Symp. Biocomput.* 474-85.
32. Azuaje F, Bodenreider O. (2004) Incorporating ontology-driven similarity knowledge into functional genomics: An exploratory study (Abstract). *IEEE Fourth Symp. Bioinformatics Bioengineering* Taichung, Taiwan, p. 317.
33. Chen Y, Xu D. (2004) Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*. *Nucl. Acids Res.* 32:6414-24.
34. Yu H, Gao L, Tu K, Guo Z. (2005) Broadly predicting specific gene functions with expression similarity and taxonomy similarity. *Gene* 352: 75-81.
35. Altman DG. (1991) *Practical Statistics for Medical Research*. Chapman & Hall, London.
36. Cox DR. (1972) Regression models and lifetables. *J. R. Stat. Soc. [B]* 34:187-220.
37. Pastor-Anglada M, Felipe A, Casado FJ, del Santo B, Mata JF, Valdes R. (1998) Nucleoside transporters and liver cell growth. *Biochem Cell Biol* 76:771-7.
38. Bode BP, Souba WW. (1999) Glutamine transport and human hepatocellular transformation. *JPN* 23:S33-7.
39. Ren Y et al. (2003) Interleukin-8 serum levels in patients with hepatocellular carcinoma: correlations with clinicopathological features and prognosis. *Clin. Cancer Res.* 9:5996-6001.
40. Poon RT, Ho JW, Tong CS, Lau C, Ng IO, Fan ST. (2004) Prognostic significance of serum vascular endothelial growth factor and endostatin in patients with hepatocellular carcinoma. *Br. J. Surg.*

- 91:1354-60.
41. Yamauchi N et al. (2005) The glypican 3 oncofetal protein is a promising diagnostic marker for hepatocellular carcinoma. *Mod. Pathol.* 18:1591-8.
  42. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. (2002) *Molecular Biology of the Cell*. Garland Publishing, New York.
  43. Kufe DW et al. (2003) *Cancer Medicine*. BC Decker, Hamilton, ON, Canada.
  44. Gascoyne RD et al. (1997) Prognostic significance of Bcl-2 protein expression and Bcl-2 gene rearrangement in diffuse aggressive non-Hodgkin's lymphoma. *Blood* 90:244-51.
  45. Pedersen LM, Jurgensen GW, Johnsen HE. (2005) Serum levels of inflammatory cytokines at diagnosis correlate to the bcl-6 and CD10 defined germinal center (GC) phenotype and bcl-2 expression in patients with diffuse large B-cell lymphoma. *Br. J. Haematol.* 128:813-9.
  46. Stros M, Ozaki T, Bacikova A, Kageyama H, Nakagawara A. (2002) HMGB1 and HMGB2 cell-specifically down-regulate the p53- and p73-dependent sequence-specific transactivation from the human Bax gene promoter. *J. Biol. Chem.* 277:7157-64.
  47. Hock BD, Haring LF, Steinkasserer A, Taylor KG, Patton WN, McKenzie JL. (2004) The soluble form of CD83 is present at elevated levels in a number of hematological malignancies. *Leuk. Res.* 28:237-41.
  48. Gibbons FD, Roth FP. (2002) Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.* 12:1574-81.
  49. Rhodes DR, Chinnaiyan AM. (2005) Integrative analysis of the cancer transcriptome. *Nat. Genet.* 37 Suppl:S31-7.
  50. Ben-Hur A, Guyon I. (2003) Detecting stable clusters using principal component analysis. *Methods Mol. Biol.* 224:159-82.
  51. Bolshakova N, Azuaje F, Cunningham P. (2005) An integrated tool for microarray data clustering and cluster validity assessment. *Bioinformatics* 21:451-5