Genomics Proteomics Bioinformatics 12 (2014) 31–38

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Pathway-based Analysis of the Hidden Genetic Heterogeneities in Cancers

Xiaolei Zhao [1,#], Shouqiang Zhong [2,#], Xiaoyu Zuo [3], Meihua Lin [1], Jiheng Qin [1], Yizhao Luan [1], Naizun Zhang [2], Yan Liang [2,*], Shaoqi Rao [1,3,*]

[1] *Institute for Medical Systems Biology and Department of Medical Statistics and Epidemiology, School of Public Health, Guangdong Medical College, Dongguan 523808, China*
[2] *Maoming People's Hospital, Maoming 525000, China*
[3] *Department of Medical Statistics and Epidemiology, School of Public Health, Sun Yat-Sen University, Guangzhou 510080, China*

Received 10 October 2013; revised 6 December 2013; accepted 9 December 2013
Available online 22 January 2014

Handled by Arndt G. Benecke

KEYWORDS

Genetic heterogeneity;
Pathway-based approach;
Sample partitioning;
Enrichment analysis;
Survival analysis;
Cancer

Abstract Many cancers apparently showing similar phenotypes are actually distinct at the molecular level, leading to very different responses to the same treatment. It has been recently demonstrated that pathway-based approaches are robust and reliable for genetic analysis of cancers. Nevertheless, it remains unclear whether such function-based approaches are useful in deciphering molecular heterogeneities in cancers. Therefore, we aimed to test this possibility in the present study. First, we used a NCI60 dataset to validate the ability of pathways to correctly partition samples. Next, we applied the proposed method to identify the hidden subtypes in diffuse large B-cell lymphoma (DLBCL). Finally, the clinical significance of the identified subtypes was verified using survival analysis. For the NCI60 dataset, we achieved highly accurate partitions that best fit the clinical cancer phenotypes. Subsequently, for a DLBCL dataset, we identified three hidden subtypes that showed very different 10-year overall survival rates (90%, 46% and 20%) and were highly significantly ($P = 0.008$) correlated with the clinical survival rate. This study demonstrated that the pathway-based approach is promising for unveiling genetic heterogeneities in complex human diseases.

Introduction

Genetic heterogeneity has attracted increasing attention in the study of genetic mechanisms of complex diseases. It describes the biological complexities that apparently similar characters may result from different genes or different genetic mechanisms [1]. In the clinical setting, patients with diseases displaying a similar phenotype but resulting from different genetic causes frequently respond very differently to the same

* Corresponding authors.
E-mail: lye30668@aliyun.com (Liang Y), raoshaoq@gdmc.edu.cn (Rao S).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.gpb.2013.12.001

treatment and thus receive a markedly different prognosis. Therefore, elucidation of the genetic heterogeneities underlying complex diseases has profound influences on both modern clinical practice and basic biomedical research.

Rapidly accumulated genomic-scale molecular data provide good opportunities to unveil the genetic heterogeneities in complex diseases at the molecular level. Significant improvements in methods and applications for analysis of the genetic heterogeneity have been achieved in the past decades. The usefulness of large-scale gene expression data, as measured by microarrays, has noticeably been indicated by the successful stratification of diffuse large B-cell lymphoma (DLBCL) [2–5]. In these pioneering studies, an unsupervised clustering algorithm was used to partition both gene expression data and patients with an aim to define genetically homogeneous novel cancer subgroups among cancer patients based on the principle that patients within the same cluster probably involve the similar molecular pathogenesis and hence could be grouped into the same molecular subphenotype [6]. Although the traditional clustering analysis based on individual gene expression profiles has achieved great success in unveiling the genetic heterogeneity, it seldom considered the combined actions of multiple functionally dependent genes. It is increasingly recognized that complex diseases such as cancers are a consequence of alterations in a complicated cascade of events involving multiple biological processes and pathways. Thus, subtypes identified by individual genes often lack good biological interpretations. In this sense, the development of function-based methods for cancer subtyping is warranted.

Gene Ontology (GO) [7] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [8] are the two most common databases currently used for gene functional annotation. GO terms are used primarily for the annotation of individual gene products, whereas KEGG pathway terms are used for the annotation of classes of gene products, thus providing a more precise delineation of functionalities for a group of genes that act together to some extent. KEGG pathway is a collection of manually drawn pathway maps that represent the knowledge on molecular interactions and reaction networks for human diseases, environmental information processing, genetic information processing, *etc*., thus possibly providing biological interpretations of higher-level systemic functions [9]. Hence, a pathway-based approach can integrate the effects of genetic factors and biological networks [10] and has been used for disease classification [11]. In our previous work [1], we proposed a GO-based approach to unveil the hidden heterogeneities in cancers, and demonstrated that it can successfully integrate the cellular function and the gene expression profile, and the approach showed the greater advantage of GO in classifying the cancer types. In principle, a similar pathway-based approach should have comparable performance in the genetic analysis of molecular heterogeneities in cancers. Numerous studies have shown that the cancer subtypes are, in essence, related to multiple pathways [12–14]. For example, recent evidence has shown that molecular subtypes of DLBCL arise from distinct genetic pathways [15]. Therefore, this study aimed to verify whether a pathway-based approach is useful in deciphering molecular heterogeneities in complex diseases such as cancer.

In this study, we proposed a pathway-based clustering approach to unveil disease heterogeneities based on multiple pathways. First, we selected differentially expressed genes that are associated with specific disease conditions. It should be noted that algorithms such as the *t* test or *F* test are not proper for selecting differentially expressed genes due to the presence of genetic heterogeneity, because the validity of these tests relies on accurately and unambiguously defining phenotype characteristics. Hence, we took a robust metric, the overall variability of gene expression, to guide gene selection. Firstly, genes with top-ranked expression variations across samples, which explain most of the total variance potentially contributed by known or unknown factors (for example, the hidden cancer subtypes), were selected as ''feature genes'' in the initial gene selection as implemented in several previous studies [16,17]. Then, we identified KEGG pathways enriched with feature genes as ''putative signature pathways'' (here, ''enriched'' means that a pathway has saliently more feature genes (with large variance) than a random gene set of the same size does). Finally, we classified samples to identify the hidden disease subtypes using the expression profiles of genes annotated to these well-characterized pathways. In the numerical analysis, we first validated the proposed approach in accurately partitioning cancer phenotypes using a publicly-available large cancer dataset. Subsequently, we used the approach to identify the hidden subtypes of a notoriously heterogeneous phenotype, DLBCL. Our results demonstrated that three new subtypes identified using signature pathways had very different 10-year overall survival rates, and the partitions were highly significantly correlated with the clinical survival rates.

Results

Validation of the proposed pathway-based approach using a large microarray dataset

We selected the signature pathways that were significantly (FDR $\leqslant$ 0.01, see the Materials and methods section for the details) enriched with the 10% top-ranked genes with largest expression variances based on the NCI60 dataset [18]. As a result, three pathways were identified, which were used for the subsequent analyses. These include the small cell lung cancer pathway (hsa05222), the extracellular matrix (ECM)–receptor interaction pathway (hsa04512) and the focal adhesion pathway (hsa04510) (**Table 1**). First, we evaluated the ability of each signature pathway to accurately partition the samples into the known cancer types using the clustering analysis based on only the expression profiles of genes within the pathway. Our results based on each of the three pathways agreed well with the original clinical labels. The observed values for the adjusted Rand index (ARI) [19] (to measure the agreement between the identified clusters and the original partitions, ranging from 0 to 1, see the Materials and methods section for the details) were 0.83, 0.69 and 0.78, respectively. Subsequently, to determine the empirical significance of each pathway, we randomly selected 1000 gene subsets of the same pathway size from the null distribution as described in the Materials and methods section. No random subset achieved an ARI value higher than that of the corresponding pathway such that all identified signature pathways showed significantly better performance ($P < 0.001$) in correctly partitioning the samples (that is, more likely relevant to the phenotypic partitions). Furthermore, after applying the majority rule voting for integrating results from the three signature pathways, we

Table 1 Signature pathways for NCI60

Signature pathway	Number of annotated genes	Nominal *P* (pathway)[a]	FDR (pathway)[b]	ARI	Number of misallocated samples	*P* (ARI)[c]
hsa05222: small cell lung cancer	19	7.83E−06	9.82E−03	0.83	2	< 0.001
hsa04512: ECM–receptor interaction	21	3.03E−07	3.81E−04	0.69	8	< 0.001
hsa04510: focal adhesion	36	1.54E−10	1.93E−07	0.78	3	< 0.001

Note: Signature pathways for NCI60 were identified by using FDR for multiple tests correction (adjusted $\alpha = 0.01$). Details of the NCI60 dataset were described previously [18]. [a] Modified Fisher Exact *P* value. [b] FDR stands for false positive rate, which is used for adjustment of multiple tests for 201 pathways. [c] Statistical significance of ARI for the selected pathway. ARI stands for adjusted Rand index.

achieved a ARI value of 0.83, with only two tumor samples misallocated. Alternatively, four samples were misclassified with construction of a decision tree (Figure 1).

We also assessed the robustness of the proposed pathway-based approach to the methods for feature gene selection. With the feature genes selected as the top 10%, 15% and 20% ranked genes with the largest variances, we found that the identified signature pathways largely overlapped. Compared to using the top 10% ranked genes as feature genes, no additional pathways were identified when using the top 15% genes, and only one more pathway was identified when using the top 20% genes. These data suggest the robustness of such pathways to the differences of the thresholds for selecting feature genes. Numerous biological experiments provided ample evidence to support the involvement of the three pathways in the molecular mechanisms underlying the various cancer types. For example, the focal adhesion pathway and the ECM–receptor interaction pathway were identified to be the functional gene sets that were significantly differentially expressed in leukemia [20]. In addition, by searching for the oncogenes in the KEGG database, one can easily find that the three pathways, particularly the small cell lung cancer pathway and the focal adhesion pathway, were enriched with various oncogenes. All evidence supports that these three pathways are truly linked to cancer(s).

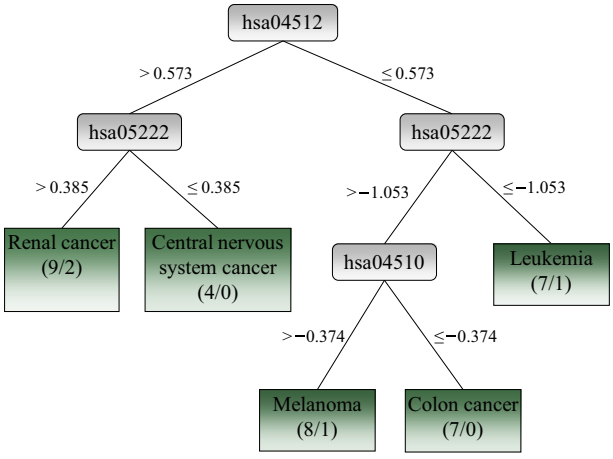

Figure 1 Decision tree based on three signature pathways for five cancer types

The internal nodes of the tree are the signature pathways. The leaf nodes represent the classification for five types of cancer (renal cancer, central nervous system cancer, melanoma, colon cancer and leukemia). Included in the leaf nodes are the total number of samples over the number of the incorrectly predicted samples for the specific type of cancer indicated.

Unveiling the hidden genetic heterogeneities in DLBCL

The genetic heterogeneities in DLBCL have been extensively investigated previously [2,21,22]. Inspired by its success in classifying the known NCI60 cancer types, we then applied the proposed pathway-based approach to discover the hidden molecular types of DLBCL.

Based on the DLBCL dataset [2], we identified three subtypes using two signature pathways, *i.e.*, the hematopoietic cell lineage pathway (has04640) and the cytokine receptor interaction pathway (has04060) (Table 2). These two pathways might be substantially responsible for the incidence and progression of DLBCL. The former pathway was associated with the immune system, and latter pathway was associated with various signaling molecules and their corresponding interactions. The abnormalities in either the immune function and/or the signaling molecules and their interactions were considered to be the major causes of DLBCL [23]. The sensitivity analysis based on different criteria for feature gene selection (top 10%, 15% and 20% ranked genes with the largest variances) revealed that the identified signature pathways largely overlapped. Compared to using the top 10% ranked genes as feature genes, two additional pathways were identified when using the top 15% genes, and only one more pathway was identified when using the top 20% genes. Thus, only the results for the criterion of the top 10% ranked genes were presented in this study.

The survival results for these subtypes are shown in Figure 2. The 10-year overall survival rates for three newly defined molecular subtypes were 90%, 46% and 20%, respectively. The log-rank statistic showed that the survival time of the three subtypes was significantly different ($P = 0.008$), which had a markedly higher caliber compared to the original partitions (the clinic labels, $P = 0.010$, see [2]) to map their differential survival profiles. Compared with the partitioning results from a GO module ($P = 0.007$) obtained previously by our group [1], the pathway-based approach performed equally well, and identified one more molecular subtype.

To further explore a compact model for clinical use, we analyzed genes included in the two signature pathways using Cox proportional-hazards models. In the univariate analysis, nine genes were found at the liberal significance level of 0.1. Subsequently, using the stepwise variable selection option (with the same inclusion and exclusion *P* values of 0.05) for the multivariate Cox proportional-hazards regression model, we identified three genes, *CD10*, *CD21* and *IL2RB*, as predictors (Table 3).

Table 2 Signature pathways for DLBCL

Signature pathway	Number of annotated genes	Nominal *P* (pathway)[a]	FDR (pathway)[b]
hsa04640: hematopoietic cell lineage	22	3.80E−10	4.76E−07
hsa04060: cytokine receptor interaction	24	1.00E−06	1.26E−03

Note: Signature pathways for DLBCL were identified by using FDR for multiple tests correction (adjusted $\alpha = 0.01$). [a] Modified Fisher Exact *P* value. [b] FDR stands for false positive rate, which is used for adjustment of multiple tests for 201 pathways. DLBCL stands for diffuse large B-cell lymphoma.

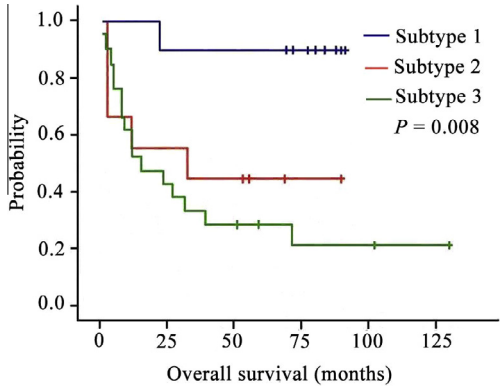

Figure 2 Clinically distinct DLBCL subtypes defined by gene expression profiling of two signature pathways

Kaplan–Meier plot of the overall survival of the three molecular subtypes of DLBCL, partitioned using the expression profiles of the genes contained in two signature pathways, hsa04640 and hsa04060.

CD10 encodes a common acute lymphocytic leukemia (ALL) antigen that serves as an important cell surface marker in the diagnosis of ALL [24]. *CD21* encodes a membrane protein that functions as a receptor for Epstein–Barr virus (EBV) binding on B and T lymphocytes. A previous study [25] reported that the prognosis of CD21-positive DLBCL was significantly favorable to that of CD21-negative DLBCL and then a later *in vivo* experiment [26] showed that CD21 was closely related to LFA-1 expression in B-cell lymphoma (BCL), and the absence of CD21/LFA-1 expression was associated with pleural/peritoneal fluid involvement caused by BCL, which is a potential indicator of BCL progression. It is interesting to note that interleukin-2 receptor beta (*IL2RB*) was significant in the Cox proportional-hazards model. Although no study has directly shown that *IL2RB* is a predictor for DLBCL, *IL2RB* has been reported to be a potential prognostic biomarker for chronic lymphocytic leukemia (CLL) [27].

Discussion

From a biological perspective, compared to GO that reflects the functional similarities of genes, KEGG pathway reflects an integration of several specific functions. It is more systematic in revealing and elucidating the sophisticated molecular mechanisms underlying complex diseases such as cancer. Several studies have suggested the link between cancer subtypes and pathways. Therefore, the proposed pathway-based clustering approach for unveiling genetic heterogeneities of complex diseases would facilitate better understanding of the mechanisms underlying these phenomena. In this study, we evaluated this approach using a public benchmark dataset. Our results demonstrated that the gene expression profiles of pathways effectively distinguished well-characterized clinical types of cancers. Hence, there was sufficient reason to believe that the putative signature pathways for a heterogeneous disease could depict the underlying molecular mechanism(s) leading to the molecular subtypes. Further application of this proposed pathway-based approach to DLBCL demonstrated its effectiveness in dissecting genetic heterogeneities in complex diseases. Similar to the GO-based approach, the proposed pathway-based approach is also an efficient unsupervised feature selection method, which yields multiple feature gene sets (*i.e.*, genes annotated to identified signature pathways) of functional compactness. The genes with top-rank expression variations across samples were selected as the initial feature genes [16,17]. Subsequently, the feature genes were further filtered or organized by significant KEGG pathways. Similar to the GO-based approach, this approach is not only useful in identifying both the gene expression signatures and the functional signatures of disease subtypes but can also provide guidance for functional studies on the molecular pathogenesis of the diseases investigated.

Although some previous studies [2] that clustered disease subtypes based on expression profiles of the genes achieved great success in dissecting genetic heterogeneities involved in DLBCL, such a clustering algorithm itself does not provide proof of the best grouping of genes in terms of biological functions [28]. Thus, biological interpretation of the grouping requires expert knowledge, which is often subjective [29]. In this study, we proposed to directly use an external annotation database such as the KEGG pathway to extract multiple functionally compact and coherent gene sets. Three hidden subtypes were identified by applying the proposed pathway-based approach in unraveling DLBCL. In terms of the survival analysis and the implications of the signature pathways, the proposed pathway-based approach provided a novel and

Table 3 Multivariate Cox proportional-hazard model built using the genes in the two signature pathways

Variable	Estimated coefficient	Wald χ^2	*P* value	Hazard ratio (95% CI)
CD10	−0.762	10.635	0.001	0.530 (0.295–0.738)
CD21	−0.735	6.210	0.013	0.467 (0.269–0.855)
IL2RB	−0.630	6.377	0.012	0.479 (0.327–0.869)

Note: CI stands for confidence interval.

feasible avenue to the genetic analysis of the hidden subtypes of complex human diseases such as cancer.

In this study, we took the known cluster number suggested by the preassigned clinic labels to validate the proposed approach, assuming the lack of heterogeneities in the NCI60 data for several well-characterized cancers. Although the clustering results provided good fits to the known phenotypic partitions, this assumption might not be true [1]. Meanwhile, the problem of estimating the correct number of clusters to unveil hidden cancer subtypes has largely remained unresolved. In addition, although the proposed pathway-based approach has achieved some success in the genetic analysis of the underlying molecular stratifications in cancer, we should recognize the limitations of this study. First, only one dataset for DLBCL was analyzed, it is thus very likely that only a small proportion of the relevant pathways were identified due to the limited information provided by a single dataset. Second, the current knowledge about pathways is largely fragmented and far from complete; hence, this limitation would compromise the aspect of this analysis that relies on pathway knowledge. Finally, although we tried our best to control type I errors (incorrect rejections of true null hypotheses) in various steps toward the identification of either pathways or hidden cancer subtypes, whether the overall type I error was well-controlled remains unclear. In this sense, we considered our analysis as exploratory in nature. Further studies using large-scale datasets and refined pathway knowledge are highly demanded, which could increase the effectiveness in detecting pathways with modest effects. Finally, although in principle the proposed pathway-based method for the analysis of genetic heterogeneities could be extended to other types of data such as that of genome-wide SNPs and next generation sequencing data, such an approach has to be carefully assessed. This assessment will be the next focus of our research group.

Materials and methods

Description of datasets

A large classical multiple-class dataset NCI60 [18] was used as the benchmark dataset to validate the efficiency of the proposed pathway-based approach, which consists of 9703 cDNAs measured in 60 cell lines of nine cancers. The data for prostate cancer were excluded because these consisted of only two samples. Samples of breast tumors, ovarian cancer and non-small cell lung carcinoma were also excluded for the possible existence of heterogeneous hidden subtypes or misassigned labeling of samples [21,30]. Thus, a subset of the NCI60 data (35 samples of five cancer types) was used in the study, including eight samples of renal cancer (RE), six samples of central nervous system cancer (CNS), eight samples of melanoma (ME), seven samples of colon cancer (CO) and six samples of leukemia (LE). After evaluating its ability for accurately partitioning this diverse data structure, we used the proposed pathway-based approach to analyze the hidden subtypes of DLBCL, which has been demonstrated to be notoriously heterogeneous [21,22,30]. The independent dataset for DLBCL consists of 4026 cDNAs measured in 42 samples [2]. We verified the identified hidden partitions by survival analysis of the clinical profiles of patients in each molecular-based partition.

A detailed procedure chart for the pathway-based approach is shown in **Figure 3** and described below. The corresponding source code is freely available upon written request. For data preprocessing, we adopted a unified criterion for the initial selection of genes from the previously described cDNA microarray datasets. First, we discarded clones with missing data in more than 5% of the arrays and applied a base-2 logarithmic transformation to the expression data. Second, similar to our previous paper [2], we imputed remaining missing data with zeros. Third, the data for genes were centered by subtracting the observed median value. The final datasets of NCI60 and DLBCL finally comprised 5124 and 3148 genes, respectively.

Selecting putative signature pathways from KEGG

For the NCI60 and DLBCL datasets, the top x percent of genes with largest expression variances were selected as feature genes. Subsequently, we loaded these feature genes into the Database for Annotation, Visualization and Integrated Discovery (DAVID) [31] software to test their enrichment in pathways based on a modified Fisher Exact test. Finally, we identified the significantly enriched pathways at a false discovery rate (FDR) of 0.01 to adjust for multiple tests of 201 pathways in the DAVID database. To demonstrate the robustness of the pathways, we compared the pathways identified at different top percentage levels (x = 10, 15 and 20) of the feature genes with the largest variances for NCI60 or DLBCL.

Clustering samples based on individual pathways

For each signature pathway, we extracted the expression profiles of the measured genes that were annotated to it. By agglomerative hierarchical clustering algorithm, each sample was initially assigned to one cluster, then the distances between 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, correlation (uncentered) was used as the distance metric and the average linkage method was used for merging. The software can be downloaded from the website of the Eisen Lab (http://rana.lbl.gov/EisenSoftware.html). We pruned off the hierarchical tree to allocate the samples into clusters. To evaluate the performance of the proposed approach, the number of clusters in the validation dataset was determined by the number of predefined clusters from the original data source. Additionally, we assessed the classification performance of these signature pathways using a decision-tree based approach [32]. Finally, the adjusted ARI [19] was calculated to measure the agreement between the identified clusters and the original partitions, which ranged from 0 to 1. The expected value of the ARI is 0 when the partitions are randomly drawn, and the ARI is 1 when two partitions perfectly agree. A larger ARI dictates a higher correspondence between two types of partitions.

To assess the significance of the ARI, we compared the observed ARI value with that of the same-sized gene subsets randomly selected from the whole microarray. The aim of this statistical test was to empirically verify whether the profiles of the genes in the signature pathways performed

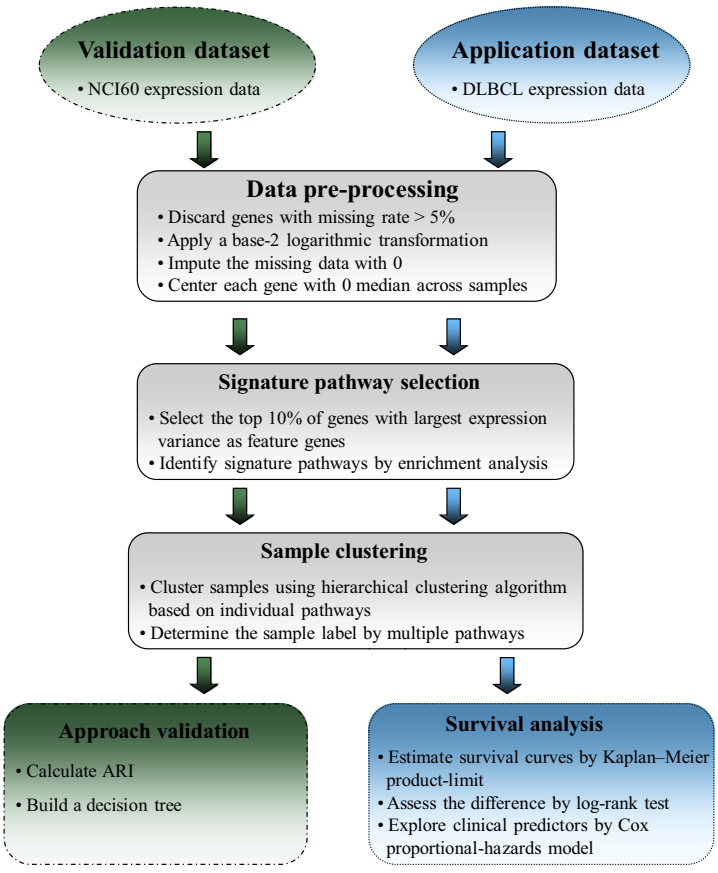

Figure 3 A detailed procedure chart for pathway-based analysis of genetic heterogeneities

significantly better at clustering than the gene groups randomly selected from a null (or contrast) population, in which the gene had no or less functional relationship. It is well known that similarly expressed (co-expressed) genes tend to share the same or similar function(s) and in fact, the gene co-expression information is often used for predicting gene functions [33]. Similar to our previous study [1], we constructed the null gene population using the silenced genes among all the annotated genes from the original expression profiles after excluding the genes annotated to the identified signature pathways and the genes significantly co-expressed with at least one gene in the signature pathways. Here, two genes were classified as co-expressed when the absolute value of Pearson correlation coefficient of their expression was larger than a threshold at significance level $\alpha = 0.005$, as determined using 10,000 gene pairs randomly sampled from the original expression profiles.

Subsequently, for each signature pathway, 1000 gene subsets of the same gene set size were randomly sampled from the null population. By applying the same clustering procedure to the 1000 random gene subsets, we defined the empirical *P* value for the observed ARI of each signature pathway as the fraction (proportion) of 1000 random subsets having ARIs larger than that of the signature pathway. The *P* value was used to assess whether an identified signature pathway had significantly better performance at correctly partitioning the samples (*i.e.*, more likely relevant to the phenotypic partitions) than the random gene subsets that were less likely to be functionally related.

Clustering based on multiple pathways

The utility of partition for a single pathway might be limited, wherein some samples could have been misclassified through the use of information from only one or a few pathways. To increase the accuracy of phenotypic partition, we applied a voting step to comprehensively integrate the partition results drawn from each signature pathway. Specifically, for a sample that had multiple membership labels obtained from different pathways, we applied a simple majority rule to determine the sample's membership. If several classes drew the vote, we randomly assigned one of the class labels to the sample.

The agreement between the clustering results based on multiple pathways and the original partitions was also evaluated by calculating the ARI. Alternatively, assuming that the original phenotypic labels for samples were correct, we evaluated the signature pathways by building a decision tree [32]. We then used the approach to unveil the hidden subtypes of DLBCL.

Survival analysis

To verify the clinical significance of the identified hidden DLBCL subtypes, we estimated the survival curves of the subtypes using the Kaplan–Meier product-limit method and assessed the difference between the survival curves using the log-rank test [34]. To explore a compact model for clinical use, we also evaluated the potential of genes in the signature

pathways for predicting phenotypes. First, we applied a univariate Cox proportional-hazards model to identify the genes whose marginal effects on the overall survival time were significant. Subsequently, a multivariate Cox proportional-hazards model was used to analyze the power of the significant genes for predicting the overall survival time. The Wald χ^2 test was used to determine the significance of each predictor's hazard toward the patients' survival time.

Authors' contributions

SR, XZ and SZ conceived the project, performed the analysis and wrote the manuscript. All the remaining authors participated in writing the computing codes and analyzing the public datasets. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 31071166 and 81373085), Natural Science Foundation of Guangdong Province, China (Grant No. 8251008901000007), Science and Technology Planning Project of Guangdong Province (Grant No. 2009A030301004), Science and Technology Project of Dongguan (Grant No. 2011108101015) and the funds from Guangdong Medical College (Grant Nos. XG1001, JB1214, XZ1105, STIF201122, M2011024 and M2011010).

References

[1] Xu JZ, Guo Z, Zhang M, Li X, Li YJ, Rao SQ. Peeling off the hidden genetic heterogeneities of cancers based on disease-relevant functional modules. Mol Med 2006;12:25–33.
[2] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503–11.
[3] Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. Proc Natl Acad Sci U S A 2000;97:12079–84.
[4] Zhang W, Li L, Li X, Jiang W, Huo J, Wang Y, et al. Unravelling the hidden heterogeneities of diffuse large B-cell lymphoma based on coupled two-way clustering. BMC Genomics 2007;8:332.
[5] Bullinger L, Dohner K, Bair E, Frohling S, Schlenk RF, Tibshirani R, et al. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. N Engl J Med 2004;350:1605–16.
[6] Bea S, Zettl A, Wright G, Salaverria I, Jehn P, Moreno V, et al. Diffuse large B-cell lymphoma subgroups have distinct genetic profiles that influence tumor biology and improve gene-expression-based survival prediction. Blood 2005;106:3183–90.
[7] Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004;32:D258–61.
[8] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 1999;27:29–34.
[9] Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res 2002;30:42–6.
[10] Chen L, Zhang L, Zhao Y, Xu L, Shang Y, Wang Q, et al. Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing SNPs and pathways. Bioinformatics 2009;25:237–42.
[11] Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol 2008;4:e1000217.
[12] Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439:353–7.
[13] Sherr CJ, McCormick F. The RB and p53 pathways in cancer. Cancer Cell 2002;2:103–12.
[14] Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med 2004;10:789–99.
[15] Lenz G, Wright GW, Emre NC, Kohlhammer H, Dave SS, Davis RE, et al. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. Proc Natl Acad Sci U S A 2008;105:13520–5.
[16] Ding CH. Unsupervised feature selection via two-way ordering in gene expression analysis. Bioinformatics 2003;19:1259–66.
[17] Dudoit S, Fridlyand J. Bagging to improve the accuracy of a clustering procedure. Bioinformatics 2003;19:1090–9.
[18] Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet 2000;24:236–44.
[19] Yeung KY, Ruzzo WL. Details of the adjusted Rand index and clustering algorithms, supplement to the paper "An empirical study on principal component analysis for clustering gene expression data". Science 2001;17:763–74.
[20] Kim J, Shin M. Identification of significant gene-sets differentially expressed in a specific disease by co-expressed functional gene modules generation. BioChip 2010;4:204–9.
[21] Monti S, Savage KJ, Kutok JL, Feuerhake F, Kurtin P, Mihm M, et al. Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. Blood 2005;105:1851–61.
[22] Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 2002;346:1937–47.
[23] Zhang ZX, Shen CF, Zou WH, Shou LH, Zhang HY, Jin WJ. Exploration of molecular mechanisms of diffuse large B-cell lymphoma development using a microarray. Asian Pac J Cancer Prev 2013;14:1731–5.
[24] Letarte M, Vera S, Tran R, Addis JB, Onizuka RJ, Quackenbush EJ, et al. Common acute lymphocytic leukemia antigen is identical to neutral endopeptidase. J Exp Med 1988;168:1247–53.
[25] Otsuka M, Yakushijin Y, Hamada M, Hato T, Yasukawa M, Fujita S. Role of CD21 antigen in diffuse large B-cell lymphoma and its clinical significance. Br J Haematol 2004;127:416–24.
[26] Tanimoto K, Yakushijin Y, Fujiwara H, Otsuka M, Ohshima K, Sugita A, et al. Clinical significance of co-expression of CD21 and LFA-1 in B-cell lymphoma. Int J Hematol 2009;89:497–507.
[27] Zhang J, Xiang Y, Ding L, Keen-Circle K, Borlawsky TB, Ozer HG, et al. Using gene co-expression network analysis to predict biomarkers for chronic lymphocytic leukemia. BMC Bioinformatics 2010;11:S5.
[28] Gibbons FD, Roth FP. Judging the quality of gene expression-based clustering methods using gene annotation. Genome Res 2002;12:1574–81.
[29] Rhodes DR, Chinnaiyan AM. Integrative analysis of the cancer transcriptome. Nat Genet 2005;37:S31–7.
[30] Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 2000;24:227–35.

[31] Dennis Jr G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 2003;4:P3.
[32] Guo Z, Zhang T, Li X, Wang Q, Xu J, Yu H, et al. Towards precise classification of cancers based on robust gene functional expression profiles. BMC Bioinformatics 2005;6:58.
[33] Obayashi T, Kinoshita K. Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. DNA Res 2009;16:249–60.
[34] Altman DG. Practical statistics for medical research. 1st ed. London: Chapman and Hall; 1991.