BMC Genomics

BioMed Central

Research article

Unravelling the hidden heterogeneities of diffuse large B-cell lymphoma based on coupled two-way clustering

Wei Zhang†[1], Li Li†[2], Xia Li*[1,2,3,4], Wei Jiang[1], Jianmin Huo[1], Yadong Wang[3], Meihua Lin[4,5] and Shaoqi Rao*[1,4,5]

Address: [1]The First Clinical College, Department of Bioinformatics, and the Bio-pharmaceutical Key Laboratory of Heilongjiang Province and State, Harbin Medical University, Harbin 150086, China, [2]Institute of Medical Genetics, Tongji University, Shanghai 200092, China, [3]Department of Computer Science, Harbin Institute of Technology, Harbin 150080, China, [4]The Biomedical Engineering Institute, Capital Medical University, Beijing 100054, China and [5]Department of Molecular Cardiology, Cleveland Clinic, Cleveland, OH 44195, USA

Email: Wei Zhang - weipoza@163.com; Li Li - flylily322@hotmail.com; Xia Li* - lixia6@yahoo.com; Wei Jiang - jiangweilh@gmail.com; Jianmin Huo - JianminHuo@ems.hrbmu.edu.cn; Yadong Wang - ydwang@hit.edu.cn; Meihua Lin - meihual410@hotmail.com; Shaoqi Rao* - raos@ccf.org

* Corresponding authors †Equal contributors

Published: 22 September 2007

BMC Genomics 2007, **8**:332 doi:10.1186/1471-2164-8-332

Received: 28 September 2006
Accepted: 22 September 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/332

Abstract

Background: It becomes increasingly clear that our current taxonomy of clinical phenotypes is mixed with molecular heterogeneity. Of vital importance for refined clinical practice and improved intervention strategies is to define the hidden molecular distinct diseases using modern large-scale genomic approaches. Microarray omics technology has provided a powerful way to dissect hidden genetic heterogeneity of complex diseases. The aim of this study was thus to develop a bioinformatics approach to seek the transcriptional features leading to the hidden subtyping of a complex clinical phenotype. The basic strategy of the proposed method was to iteratively partition in two ways sample and feature space with super-paramagnetic clustering technique and to seek for hard and robust gene clusters that lead to a natural partition of disease samples and that have the highest functionally conceptual consensus evaluated with Gene Ontology.

Results: We applied the proposed method to two publicly available microarray datasets of diffuse large B-cell lymphoma (DLBCL), a notoriously heterogeneous phenotype. A feature subset of 30 genes (38 probes) derived from analysis of the first dataset consisting of 4026 genes and 42 DLBCL samples identified three categories of patients with very different five-year overall survival rates (70.59%, 44.44% and 14.29% respectively; $p = 0.0017$). Analysis of the second dataset consisting of 7129 genes and 58 DLBCL samples revealed a feature subset of 13 genes (16 probes) that not only replicated the findings of the important DLBCL genes (e.g. *JAW1* and *BCL7A*), but also identified three clinically similar subtypes (with 5-year overall survival rates of 63.13%, 34.92% and 15.38% respectively; $p = 0.0009$) to those identified in the first dataset. Finally, we built a multivariate Cox proportional-hazards prediction model for each feature subset and defined *JAW1* as one of the most significant predictor ($p = 0.005$ and 0.014; hazard ratios = 0.02 and 0.03, respectively for two datasets) for both DLBCL cohorts under study.

Conclusion: Our results showed that the proposed algorithm is a promising computational strategy for peeling off the hidden genetic heterogeneity based on transcriptionally profiling disease samples, which may lead to an improved diagnosis and treatment of cancers.

Background

When a patient is diagnosed with cancer, various clinical parameters are used to assess the patient's risk profile. However, the patients with a similar prognosis frequently respond very differently to the same treatment. This may occur because two apparently similar tumours are actually completely different diseases at the molecular level, often called genetic heterogeneity. It describes the biological complexity whereby apparently similar inheritable characters result from different genes or different genetic mechanisms. The presence of such heterogeneity has a significant impact on both the efficiency of modern clinical practice and biomedical research of common human diseases. Gene chip technology measuring the transcriptional omics holds a promise in tackling the heterogeneity issues for complex human diseases, i.e., the subtypes of a disease can be discovered accurately at a molecular level by analysis of the gene expression profiles. Recent examples can be witnessed in the studies of leukaemia [1,2], breast cancer [3,4], renal allograft [5], lung cancer [6,7] and prostate cancer [8], based on unsupervised hierarchical clustering. Diffuse large B-cell lymphoma (DLBCL) analyzed in this study is the most common type of lymphoma in adults and demonstrates very apparently clinical heterogeneity. It can be treated by chemotherapy in only approximately 40% of patients. Several recent studies used DNA microarrays to study DLBCL, suggesting that it is possible to identify subgroups of patients in terms of different survival courses via gene expression data [9,10], which are unlikely to be discovered by traditional clinical approaches.

However, most of the methods for peeling off heterogeneities resort to the unsupervised learning techniques, such as hierarchical clustering, to identify clinically relevant subtypes based on all genes or a large number of genes on microarrays. Their utility is limited when the disease heterogeneity is resulted from only a small subset of the genes that participate in a particular cellular process, leading to different clinical outcomes. When the full dataset is analyzed, the "signal" of this process may be completely overwhelmed by the "noise" generated by the vast majority of unrelated data. In this study, we thus proposed an improved heterogeneity analysis strategy over the coupled two-way clustering algorithms [11-13]. In the proposed two-way clustering algorithm, super-paramagnetic clustering (SPC) algorithm [13,14] was used to take its advantages as an efficient partitioner: the number of clusters was achieved by the algorithm internally, without a need to be externally prescribed; and its stability against noise, thus providing a mechanism to identify robust stable phenotypic clusters using the most compacted subset(s) of gene signatures that leads to the best fits of the sample partitions. The rapidly accumulated multiple lines of evidence from, among others, gene expression and protein-protein interaction studies, support that genes express and perform their highly integrated cellular functions in modular fashions in cells [15-17]. Also inspired by our recent success in peeling off the hidden genetic heterogeneities of cancers based on disease relevant functional modules [18], we further defined a GeneOntology (GO)-based [19-21] conceptual functional similarity measure in order to establish a functional validation for the identified gene subsets. Finally we proved the differential survival outcomes of new subtypes using Kaplan-Meier survival analysis and multivariate Cox proportional-hazards prediction modelling according to their clinical data. We demonstrated the behaviours and properties of the proposed method by applying it to two publicly available microarray datasets of diffuse large B-cell lymphoma (DLBCL), a notoriously heterogeneous phenotype.

Results

Description of DLBCL datasets

In this study, we used two published gene expression data for DLBCL. The first dataset, analyzed initially by Alizadeh et al. [9], consists of 42 samples, and 40 of them have survival data as well. The microarray data, available at Lymphoma/Leukemia Molecular Profiling Project [22], the website companion to [9], have expression profiles for 4026 genes, and among the 4026 genes, 1980 genes have missing values. We imputed missing values by the *K*-Nearest Neighbours method ($K = 5$) [23]. The second dataset, analyzed initially by Shipp et al [24] and available at the website for The Broad Institute's Cancer Program Data Sets [25], consists of 58 samples and 7129 genes.

Coupled two-way clustering

We searched significant gene subsets using the well established coupled two-way clustering (CTWC) algorithm as implemented in a public sever [26], which used SPC as the underlying clustering tool to break down the total dataset into subsets of genes and samples iteratively until significant partitions (submatrices) were revealed. First, we clustered all samples using all genes to identify stable sample partitions and clustered all genes using all samples to identify stable gene subsets. Then, we clustered the genes gained in the previous step using the newly defined sample partitions (including all samples) to find the responsible gene subsets of high discriminating power. Finally, we clustered each sample partition again using each gene subset with high discriminating power. In the searching process, we explored the cluster depth for both dimensions of samples and genes. The cluster depth selected was based on the empirical judgement whether the clinical samples could be well separated using the candidate gene subset(s). For the dataset of Alizadeh et al [9], we stopped the sample clustering at the cluster depth of one, with eight stable and significant gene subsets (G_2, G_3, ..., G_9) identi-

fied. For the dataset of Shipp et al [24], we also stopped the sample clustering at the cluster depth of one, with 30 stable and significant gene subsets (G_2, G_3, ..., G_{31}) identified.

Computing the functional concept consistency scores for the identified gene subsets

Co-expression genes often share some functional relevance. Because we clustered the samples based on expression similarity (and difference) among the putative feature genes, the samples within a cluster were expected to be more similar in transcriptional activities than those in different clusters. Hence, the sample partitions might reflect their differences in response to the underlying biology pathway(s) leading to the phenotypic differentiation. In order to establish the functional validation for the identified gene subsets, we defined a GeneOntology (GO)-based [19-21] conceptual functional similarity measure, called concept consistency score (see Methods for detail).

The CTWC algorithm was able to identify numerous highly correlated gene subsets during the recursive partitioning of samples and genes. In this study, our goal was to find some partition of DLBCL with high medical implications. The consistency scores for eight stable gene subsets (G_2, G_3, ..., G_9) identified in the dataset of Alizadeh et al [9] were 1.00, 0.00, 0.11, 0.46, 0.30, 0.00, ∅ and 0.00, respectively. An empty value ∅ occurred because some genes had neither functional annotation nor a common parent node in GO. Subset G_2 had the highest consensus score, and was thus selected as a modular signature for subtyping the disease samples. The consistency score of G_4 was the highest (score = 0.48) among 30 stable gene subsets (G_2, G_3,, G_{31}) identified in the dataset of Shipp et al [24]. Their scores were 0.44, 0.25, 0.48, 0.00, 0.41, 0.00, 0.39, 0.10, ∅, 0.38, 0.37, 0.00, 0.33, 0.32, 0.11, 0.05, 0.28, 0.27, 0.42, 0.24, 0.20, 0.19, 0.11, 0.01, 0.16, 0.15, 0.00, 0.11, 0.1, 0.31, respectively.

Clustering samples by SPC, using the gene subsets with highest score as the modular features

When clustering a dataset using a subset of genes, it is important to know if the samples can be well characterized using such a subset. For the dataset of Alizadeh et al, forty-two DLBCL samples were clustered using the genes included in G_2 using SPC, with the Euclidean distance and Pearson's correlation coefficient as the sample and the gene expression similarity measures, respectively. Figure 1, plotted by Treeview [27,28], shows clearly a partition of three subtypes of diffuse large-B-cell lymphomas for 42 patients per the expression patterns of 38 probes (representing a total of 30 unique genes/loci). Among 38 probes, 27 were annotated in GO: 6 were for the major lymphoid-restricted membrane protein (*JAW1*), 2 for the B-cell CLL/lymphoma 7A (*BCL7A*), 2 for phosphoinositide-3-kinase, catalytic, gamma polypeptide (*PIK3CG*) and 2 for cancer susceptibility candidate 1 (*CASC1*). In addition, *BCL7A* and *TNFA* were previously reported as the prognostic factors for lymphoma. Five of the remaining 11 probes were known transcribed loci. A careful scrutiny of G_2 revealed that it largely captured complex modular activities that regulate cell cycling, DNA synthesis and repair, leukocyte adhesion, cell-cell signalling etc. and that mainly take place in nucleus and intracellularly. It is also interesting to note that G_2 contained an "integral to plasma membrane" component consisting of the well known DLBCL relevant pathway – G-protein coupled receptor protein signalling pathway (e.g. *LANCL1*, *CASC1* and *PIK3CG*) [29] and *JAW1* for vesicle targeting and homocyte development. For the functional annotations for the known genes included in G_2, see Additional File 1.

For the second DLBCL dataset, 58 DLBCL samples were clustered using the genes included in G_4 using SPC. We again identified a partition of three subtypes of DLBCL among 58 patients per the expression patterns of 16 probes, as shown in Figure 2. Among those, 15 probes were annotated in GO: 3 were for the major lymphoid-restricted membrane protein (*JAW1*), 2 for the B-cell CLL/lymphoma 6 (*BCL6*), 1 for the B-cell CLL/lymphoma 7A (*BCL7A*), 1 for the Cylin D2. Most of the 15 genes were found to be germinal centre B cell signatures [24], suggesting that G_4 described a complex process leading to a favourable survival outcome for DLBCL patients. However, we recognized that compared with G_2 some new genes (or functions) were identified in G_4, which may define some new pathways or expand our knowledge on the functional topology for DLBCL, or may be simply due to the differences between the two datasets. For the functional annotations for the known genes included in G_4, see Additional File 2.

It is interesting to note that several genes were repeatedly identified as important molecular signatures for DLBCL. For example, multiple probes for *JAW1* were repeatedly identified in both datasets. In addition, many previous experiments also detected the overexpression of this gene in normal [30] or impaired germinal centre (GC) B-cells [31] and *JAW1* had thus been established to be one of the most important molecular signatures for the GCB subtype of DLBCL [9,32-34]. The *JAW1* gene, encoding a lymphoid-restricted membrane protein (hence also called LRMP), was first identified by screening for genes expressed preferentially in B-cell lines [35] and later found in lymphoid tissues and in the pancreas and colon [36]. A recent immunohistological study of B-cell lymphomas [31] documented *JAW1* expression at the protein level in human tissues by using immunohistochemical and western blotting. And the investigators found that *JAW1*-

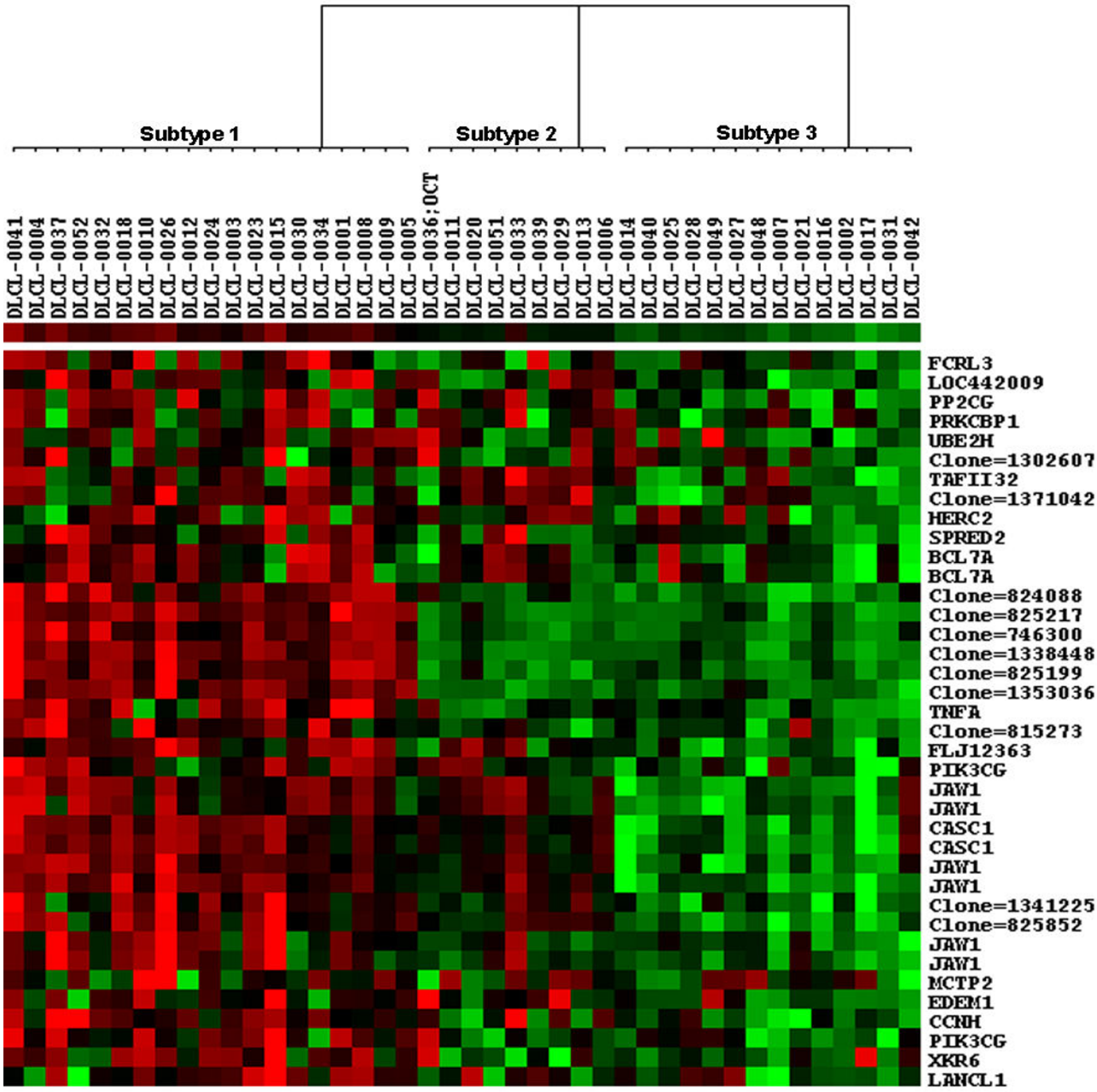

Figure 1

The three partitions of DLBCL were identified using G_2 as the disease feature set in the Alizadeh et al's dataset. In the figure, each gene corresponds to a row, and each DLBCL sample corresponds to column. Forty-two DLBCL samples were divided into three subtypes (Subtype 1, Subtype 2 and Subtype 3). Red areas indicate increased expression, and green areas decreased expression. Genes that are characteristically expressed in three subtypes of diffuse large-B-cell lymphomas are indicated. The dendrogram at the top shows the degree to which each DLBCL subtype is related to the others with respect to gene expression.

encoded protein was also highly expressed in germinal centre B-cells. Overall, multiple lines of evidence at different molecular levels support that *JAW1* is the most important prognostic marker for lymphoma. Interestingly, the *JAW1* gene has been reported to be fused to the *BLC6* gene in a case of transformed follicle centre lymphoma [37]. In this study, both genes were included in G_4 for the second dataset.

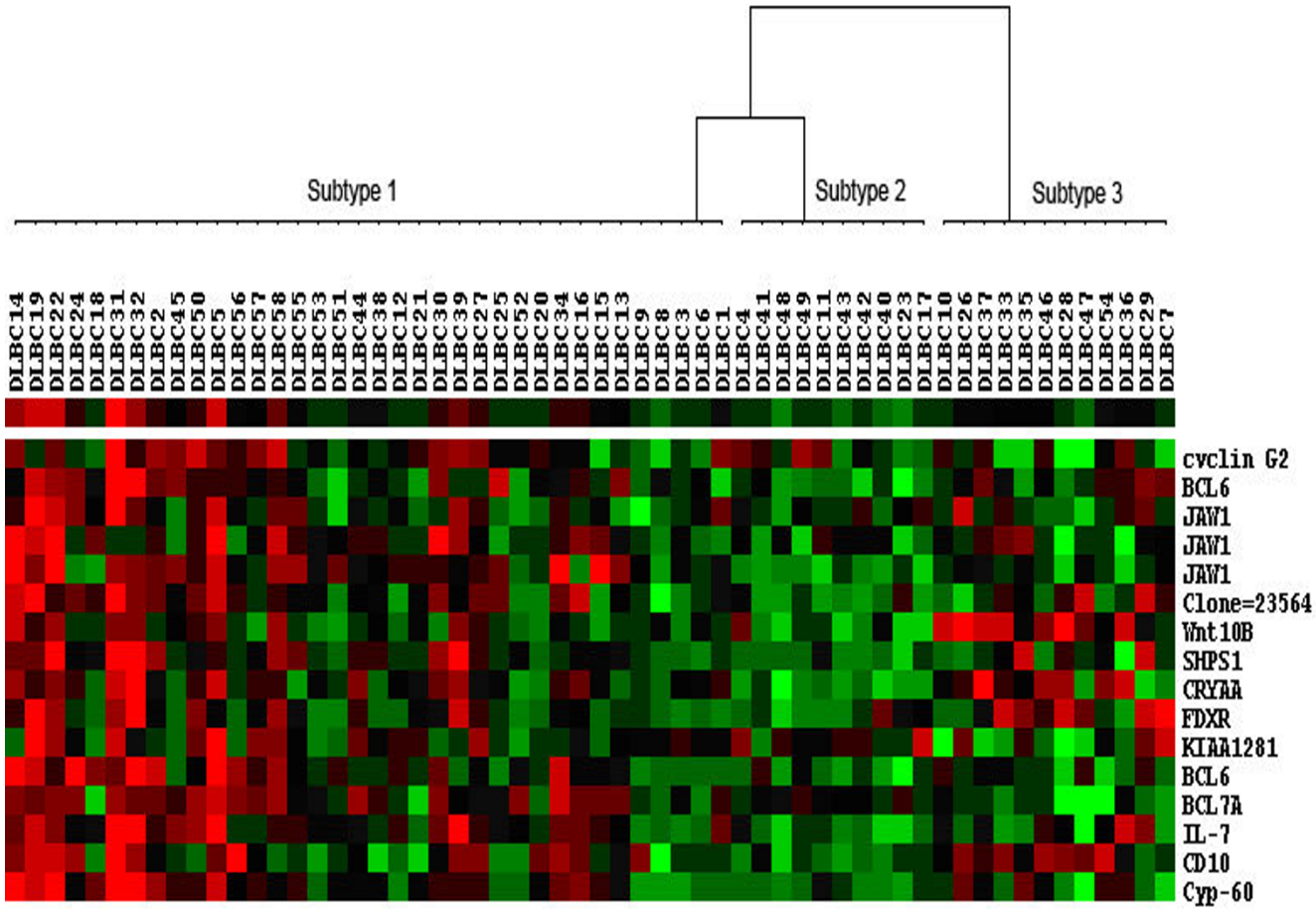

Figure 2
The three partitions of DLBCL were identified using G_4 as the disease feature set in the Shipp et al's dataset. In the figure, 58 DLBCL samples were divided into three subtypes (Subtype 1, Subtype 2 and Subtype 3).

BCL6, *BCL7A*, *KIAA1281*, *Cyclin G2*, *PIK3CG* and *JAW1* included in the subsets of G2 and/or G4 for the two datasets are previously reported germinal centre-associated signatures [24]. *BCL6* was the prognostic marker for GCB-like DLBCL and non- GCB-like DLBCL groups and *JAW1* has different expressions among the three subtypes identified by Wright et al. [32]. Consequently, it is not surprising that both subsets of genes had high discriminating power in recognizing GCB subtype of DLBCL. Based on the definitions of DLBCL subtypes proposed by Alizadeh et al. [9], most of subtype 1 partitioned by G_2 were of germinal centre B-cell-like (GCB-like) DLBCL (17 GCB-like DLBCL cases: 2 activated B-like (AB-like) DLBCL cases), and all of subtype 3 were of AB-like DLBCL. However, the clinicopathological characteristics of subtype 2, consisting of 5 AB-like DLBCL cases and 4 GCB-like cases were less clear. As shown in Figure 1, the subtype 2 defined by G_2, which may correspond to Rosenwald et al's type 3 DLBCL [10], did not express the set of genes of G_2 at a high level.

Survival analysis

To verify the clinical significance of the identified hidden DLBCL subtypes, we estimated survival curves by using Kaplan-Meier product-limit method and assessed the differences between the survival curves of the subtypes of DLBCL patients by a log-rank test [38]. For the first dataset, the survival curves associated with the three subtypes revealed by G_2 are shown in Figure 3. The log-rank statistic comparing the survival times of the first subtype and the third subtype (as shown in Figure 1) shows highly significant differences ($p = 0.0017$). The 5 year survival rates for three subtypes were 70.59%, 44.44% and 14.29%, respectively. The survival curves associated with the three subtypes identified by G_4 in the second dataset are similar to the plots for the first dataset except for subtype 3 that has a steeper survival curve (Figure 4). The log-rank statistic comparing the survival times of the first subtype and the third subtype shows highly significant differences (p =

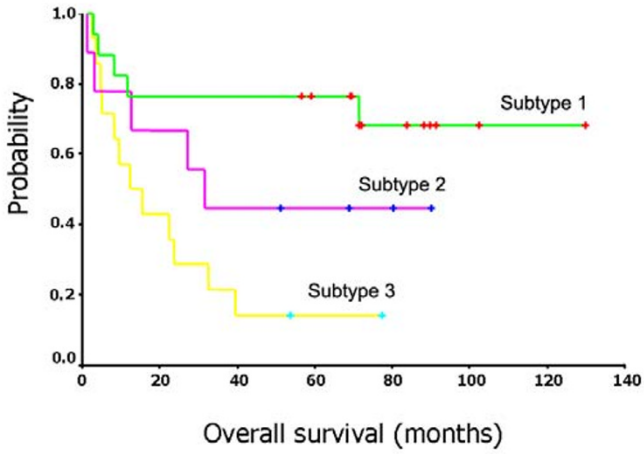

Figure 3
Survival curves for three subtypes of the DLBCL patients in the Alizadeh et al's dataset.

0.0009). The 5 year survival rates for three subtypes were 63.13%, 34.92% and 15.38%, respectively.

In order to explore a compact model for clinical use, we further identified the most contributed genes of high prediction power. Multivariate Cox proportional-hazards model was used to analyze the genes in G_2 and G_4, respectively. To reduce the number of variables to be modelled, we applied the stepwise variable selection option (with the same inclusion and exclusion p value of 0.05) for the multivariate Cox proportional-hazards regression model [39]. We ended up with four predictors (genes) for G_2 and two predictors for G_4, respectively (Tables 1 and 2). Not surprisingly, both *BCL6* and *JAW1* were selected to be the significant prognostic predictors because of their importance involved in the underlying pathogenic mechanisms for lymphoma. Hans et al [40] estimated that it was feasible to assign patients to GCB-like and non-GCB-like groups based on only three markers (*CD10*, *BCL6*, and *MUM1/IRF4*). However, the finding of *JAW1* being repeatedly selected for both datasets may implicates its discriminating role in lymphoma patients at large. In both datasets, the *JAW1* gene was defined as an important 'favourable' prognosis predictor (p = 0.005 and 0.014; hazard ratios = 0.02 and 0.03, respectively for two datasets). Based on the GO function data, all other three genes (*PRKCBP1*, *TAFII32*, *CCNH*) in G_2 are associated with the functions of 'regulation of transcription', implying their putative roles in cell development and cycling (also see Additional File 1).

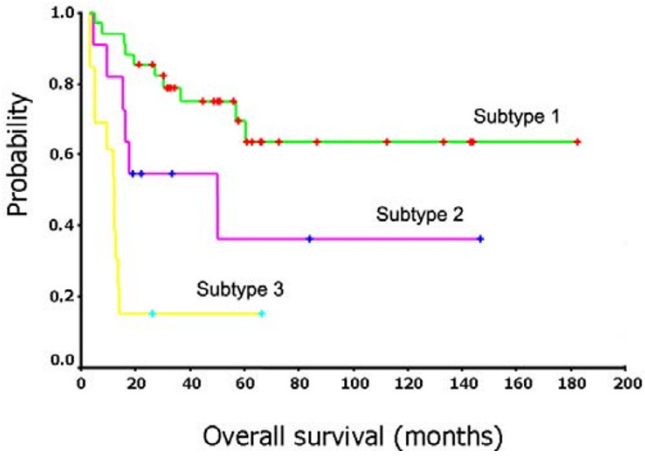

Figure 4
Survival curves for three subtypes of the DLBCL patients in the Shipp et al's dataset.

Discussion and conclusion

An initial gene expression profiling study of DLBCL led to the discovery that this single diagnostic category consists of at least two molecularly distinct diseases [9]. One DLBCL subtype, termed GCB-like DLBCL, expressed genes characteristic of normal germinal centre B cells whereas the other subtype, termed AB-like DLBCL, instead expressed genes characteristic of mitogenically activated blood B cells. Patients with GCB-like DLBCL were more often cured by chemotherapy than patients with AB-like DLBCL were. Recently, in an expanded gene expression profiling study of 274 DLBCL patients, the two gene expression subtypes were again identified together with a new subtype, termed type 3, which did not express the genes characteristic of either GCB- or AB-like DLBCL [10]. As before, patients with GCB-like DLBCL had a more favourable clinical course, with a 5-yr survival rate of 60% compared with 5-yr survival rates of 35% and 38% for patients with AB-like and type 3 DLBCL, respectively. Another study used oligonucleotide microarrays to profile gene expression in 58 DLBCL biopsies [24] and attempted to identify the GCB- and AB-like DLBCL subtypes using the genes that were identified in the original profiling study for distinguishing these subtypes [9]. Hierarchical clustering of the DLBCL cases based on expressions of these genes resulted in two groups of patients that did not differ in clinical outcome [24], in apparent contrast with the two other studies [9,10]. Compared with the previous studies, the two-way clustering algorithm applied in this study appears more efficient in finding the most compact gene subsets that have achieved an improved prognostic accuracy over the DLBCL patients' survival profiles. However, it should be cautioned that more detailed clinicopathologic characteristics for subtype 2 DLBCL patients as defined by either gene subset G_2 or G_4 have to be fully characterized before use although the survival profiles for the subtype can be clearly separated from other two subtypes. Based on the gene expression patterns of G_2 (Figure 1), it appears that many genes in subtype 2 patients were

Table 1: Multivariate Cox proportional-hazards analysis based on the G_2 signature genes relevant to survival time

Variable	Estimated coefficient	Wald χ^2	*p* value	Hazard ratio (95% CI)
PRKCBP1	4.16	9.08	0.003	63.97 (4.28–956.77)
TAFII32	4.07	9.16	0.002	58.55 (4.20–817.04)
CCNH	4.23	10.40	0.001	68.55 (5.25–895.07)
JAW1	-3.86	7.99	0.005	0.02 (0.00–0.16)

inactivated and hence the G_2's ability in differentiating these samples was significantly lowered.

Computational discoveries of the hidden subtypes for a complex disease have to be verified by some means, e.g., a functional assay using bioinformatics approaches or a clinical validation using epidemiological approaches such as survival analysis. In supervised classification, the choice of the best subset of genes for disease prediction should be relatively easy because the sample labels in training set are given, the high accuracy rate(s) of the classifiers trained on the candidate subsets might be used to filter more specific and critical subsets highly relevant to a disease pathogenesis. In unsupervised clustering analysis, however, identifying the best subset for peeling clinically heterogeneous disease can be a very challenging task as no cross-validation can be done internally. The underlying assumption for a clustering algorithm is that genes with similar expression patterns are more likely to have a similar biological function(s), but a clustering algorithm itself does not provide proof of the best grouping of genes in terms of biological functions [41]. Thus, the biological interpretation of the disease clustering results relies heavily on the expert knowledge which often is somewhat subjective [42]. Therefore, in this study, we designed a functional consensus score for evaluating a candidate gene subset in terms of functional concept consistency, which is similar to the biological homogeneity index (BHI) proposed recently [43]. Based on evaluating the performance of ten well-known clustering algorithms on two gene expression datasets, the authors in [43] found that a good clustering algorithm should have a high BHI. Alternatively, one can use the external annotation database such as Gene Ontology to directly guide the selection of multiple functionally compact and coherent gene subsets (modules) as we did in a recent study [18]. In terms of the better-characterized functionality of subsets G_2 and G_4 and based on the significantly different survival results for the patients defined by the newly defined subtypes, the applied two-way clustering algorithm has been demonstrated to be a feasible and promising toolbox for peeling off molecular heterogeneities of complex human diseases.

In this study, we took the known subtypes suggested by previous studies as the basis to assess the validity of the proposed approach. Although the clustering results provided good fits to these known phenotypic partitions, the implied assumption of the lack of other subtypes or subtle DLBCL groups might not be true. Also, the problem to estimate the correct number of subtypes for peeling off complex diseases is not investigated in this study. Some investigators [44-46] have proposed several methods to obtain the best number of sample partitions by optimizing some validity indices such as the adjusted Rand index (ARI) [44], which would provide additional insights onto improving the two-way clustering algorithm applied in this study.

There is a growing interest in biomedical domains for developing robust predictive model for the survival of cancer patients using gene expression data. However, many methods use all the genes on chips or a large number of genes (e.g. those filtered according to a marginal threshold) to predict a survival. Since the vast majority of the genes in a given dataset are irrelevant to the survivals of the studied patients, the result is that many of the inputs to the predictive model are superfluous and thus reduce the accuracy of the model for prediction. Hence, McLachlan et al. [47] proposed a mixture model-based approach to the clustering of microarray expression data. In this approach, a subset of the genes relevant for the clustering of the tissue samples was first selected by fitting mixtures of t distributions to rank the genes in order of increasing size of the likelihood ratio statistic for the test of one versus two components in the mixture model. Then, if this reduced set of genes is still too large for a normal mixture model to be fitted directly to the tissues, the investigators suggested the use of mixtures of factor ana-

Table 2: Multivariate Cox proportional-hazards analysis based on the G_4 signature genes relevant to survival time

Variable	Estimated coefficient	Wald χ^2	*p* value	Hazard ratio (95% CI)
BCL6	-3.42	4.58	0.032	0.20 (0.01–0.72)
JAW1	-3.03	6.00	0.014	0.03 (0.00–0.48)

lyzers to reduce the dimension of the feature space of genes further. In this study, we applied an integrative approach that combines a SPC-based two-way clustering with a functional consensus metric to identify functionally sounding and the most compact subset of genes underlying the phenotypic partitions of patients. Application of the proposed approach to two DLBCL datasets led to identification of two gene subsets with several features overlapped, and further multivariate Cox proportional-hazards modelling defined *JAW1* as one of the most significant predictors for the survival of the DLBCL patients in both cohorts. Overall, our results demonstrated that the proposed approach is promising for peeling off the hidden genetic heterogeneity based on modern omics data, and may lead to an improved diagnosis and treatment of cancers.

Methods

Super-paramagnetic clustering

SPC is a newly developed clustering method by mimicking the physical attributes of inhomogeneous ferromagnets. A detailed description of the algorithm can be found in [14,48]. Here only a brief introduction to the method is provided. At first, SPC builds a weighted graph for a putative data partition by computing the linkage edge weights of each object and its K nearest neighbours, respectively. Then, it evaluates each data partition using a cost function. Finally, it identifies each cluster through combining each kind of partitions.

(1) Weighted graph

For each clustering object data Z_i (i = 1, ..., N), a feature vector that corresponds to a point in a D-dimensional space, we computed the distance: $d_{ij} = |Z_i - Z_j|$, (i, j = 1, ..., N). If Z_j was one of the K closest neighbours of Z_i, then we connected the two points Z_i and Z_j by an edge with a weight:

$$J_{ij} = J_{ji} = \frac{1}{K}\exp\left(-\frac{\left\|Z_i - Z_j\right\|^2}{2a^2}\right),$$

where α was the average of d_{ij}, and K was the number of neighbours for an object. We fixed K = 10.

(2) Cost function for graph partitions

We randomly assigned an integer label L_i = 1, 2, ..., q $\{L_1, L_2, ..., L_N\}$ to the i-th object to produce a partition $\{Z\}$. If $L_i = L_j$ in a partition then Z_i and Z_j belong to the same cluster C. Otherwise, they were in different clusters. For the simulation, we fixed q = 20.

The cost function of $\{Z\}$ was:

$$H(Z) = \sum_{\langle i,j \rangle} J_{ij}(1 - \delta_{L_i,L_j}),$$

where

$$\delta_{L_i,L_j} = \begin{cases} 1 & \text{if } L_i = L_j \\ 0 & \text{otherwise, i.e. } L_i \neq L_j \end{cases}.$$

The lowest cost $H(Z)$ = 0 was obtained when all data points belong to one group; the highest cost was reached if none of its neighbours was from the same group. The smaller distance between two points relates a higher likelihood that they belong to the same group. Hence the value of $H(Z)$ reflects the resolution at which the partition $\{Z\}$ views the data.

(3) Ensemble of partitions

We considered all configurations $\{Z\}$ that had (nearly) the same value of $H(Z) = E$ rather than choosing any particular partition (say by minimizing the cost function). In the resulting statistical ensemble of partitions, each $\{Z\}$ appeared with the statistical weight $P(\{Z\}) \propto e^{-H(z)/T}$: at T = 0 only groupings with E = 0 had a non-vanishing weight; at $T = \infty$ all partitions had an equal weight. For a sequence of values of the temperature T, we calculated, by Monte Carlo simulation, the average of $R_{ij} = \left\langle \delta_{L_i,L_j} \right\rangle$, the probability of Z_i and Z_j in the same cluster at the resolution set by T.

(4) Identifying clusters

The "stable" clusters were discovered under the null hypothesis specified by R_{ij} by following a three-step procedure. First, we built the cluster "core" using threshold R_{ij}. For every pair of neighbours Z_i and Z_j, if $R_{ij} > 0.5$, we set a "link" between Z_i and Z_j. Second, we captured the points lying on the periphery of the clusters by linking each point Z_i to its neighbour Z_j of the maximal correlation R_{ij}. Third, we identified the data clusters from the linked components of the graphs obtained in the former two steps.

In the SPC procedure, a tuneable parameter T ("temperature") controlled the resolution of the performed clustering. One started at T = 0, all the objects dropped in a single cluster. As T increased, this cluster broke into several subclusters that reflected the structure of the data. Clusters kept breaking up as T was further increased, until each object formed its own cluster at high enough values of T_{max}. At last, SPC formed a hierarchical dendrogram.

As opposed to most agglomerative algorithms, SPC had a natural measure of relative stability over a range of temperatures, ΔT_c, in which the cluster retained unchanged.

The more stable cluster was expected to "survive" over a larger range of ΔT_c. For evaluating the stability of a cluster, we set a value for ΔT_c, above which a cluster was considered as stable. In order to obtain ΔT_c, we randomly permuted elements of the expression matrix under investigation, and applied SPC to the randomized matrix. ΔT_c was determined until no clusters satisfied $\Delta T_c > \Delta T$ among 500 different random permutations. This gave a bound on the probability that the clusters that we labelled as stable were in fact an artefact of noisy data. For a stable cluster, the larger the range ΔT_c was, the more stable it was. Otherwise, if the number of objects involved in a stable cluster was small, SPC considered it as a noisy cluster. Generally, the value is set to be five.

The CTWC method

The applied CTWC algorithm is a heuristic and iterative method [13]. For a gene expression profile matrix M, we denoted the initial sample set as S_1, and the gene set G_1. Clustering gene set G_i on the basis of their expression levels over the set of samples S_j was referred to the process in an operation denoted by $G_i(S_j)$. Similarly defined, $S_j(G_i)$ described the process in clustering S_j using all genes of G_i. The computational procedures for the CTWC method can be described as follows:

(1) Initialization

Compute $S_1(G_1) = \{S_j\}$, (j = 2, 3, ...), and then $G_1(S_1) = \{G_i\}$, (i = 2, 3, ...). Now the cluster depth equals to 0.

(2) Identification of stable clusters of genes and samples

Find the most stable G_i (i = 2, 3, ...) and S_j (j = 2, 3, ...) per the stability described previously. Compute $S_j(G_i)$ (including S_1) and $G_i(S_j)$ (including G_1) for clusters of depth of 1.

(3) Iterations

Repeat (2) until the updated clusters were smaller than some fixed threshold or the maximally allowed cluster depth was reached.

Evaluation of a gene subset by functional concept consistency using GO

GO describes functions of genes and relationships between genes using standard terms. It annotates the functions of a gene from dimensions of molecular function, biological process and cellular component. Generally, genes that take part in a same biological process (such as a metabolism pathway or signal transduction pathway) or that are situated in a proximate subcellular location, often share some function(s) [18].

We obtained many high-correlation sample subsets and gene subsets. In order to identify tumour subtypes both biologically meaningful and clinically relevant, we evaluated a functional concept consistency score for a gene subset to define its biological meanings. The GO-based consistency score was proposed to measure the functional consensus of the entire set of clusters produced by some unsupervised clustering algorithms such as super-parametric two-way clustering used in this study. The aim for developing this metric was to expand current clustering algorithms to produce biological meaningful clusters that are not only able to find the stable partition(s) hidden in the data, but also are useful for elucidating the underlying mechanisms leading to distinct molecular forms in phenotypically defined disease. This index is similar to the biological homogeneity index recently proposed by Datta and Datta [43], measuring how biological homogeneous the clusters are. It may also be considered as a broader and across-GO-node function definition of a cluster of genes, and is particularly useful for evaluating a gene subset that have more than one common function and an included gene is annotated with multiple classes of functions.

The steps for computing a concept consistency score [49] for gene subsets G_i (i = 1, 2, ..., N) were described as follows:

(1) Functional annotation

For $g_j \in G_i$, we mapped gene g_j to a node(s) of GO via the links between three databases: GeneBank, Unigene, and LocusLink. As a result, we obtained a concept set ($U_j = \{e_{j1}, e_{j2}, ..., e_{jm}\}$) for gene g_j.

(2) Distance between concepts

Also, for $g_k \in G_i$, the concept set of g_k, $U_k = \{e_{k1}, e_{k2}, ..., e_{kn}\}$ was obtained. We computed the distance between g_j and g_k:

$$d(U_j, U_k) = \underset{h \le |U_j|, l \le |U_k|}{\arg\min} \left\{ d(e_{jh}, e_{kl}) \right\},$$

where

$$d(e_{jk}, e_{kl}) = \text{depth}(e_{jk}) + \text{depth}\ (e_{kl}) - 2\text{depth}\ d(e_{jk}, e_{kl})$$

depth(e) denoted the depth of concept e, say the distance between e and the root of GO, and depth(e_{jh}, e_{kl}) denotes the depth of the nearest common father of the concepts e_{jh} and e_{kl}.

(3) Concept consistency

For each G_i,

$$CC(G_i) = 1 - \underset{i, j \le |G_i|, i \ne j}{\arg\text{mean}} \left\{ d(U_i, U_j) \right\}.$$

A higher $CC(G_i)$ corresponds to a higher degree of functional consistency among the genes involved in G_i. We

knew that the deeper the node of GO was, the more specific the function description was. Ideally, we should consider the hierarchical structure of GO when we computed concept consistency between two nodes. In this study, we added a weight to each concept to improve the similarity estimates between the nodes. The weight of each concept was: $W = W_0^d$, where d = depth(e). The smaller the weight was, the deeper the concept was. The value of W_0 was assigned to be $W_0 = 0.75$ in this study.

Survival analysis

Prior to survival analysis, we obtained the arithmetic mean from the data of multiple probes that correspond to an identical gene. 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 a log-rank test [50]. To construct a model for predicting the overall survival time, a multivariate Cox proportion-hazards model [39] was used to determine the significance (at significant level $p < 0.05$) of the effects of the genes included in the identified gene subset(s) on the patients' survival months. Wald Chi-square test was used to determine the significance of each predictor's hazard toward the survival time.

Computational algorithms

The algorithm flow for the proposed heterogeneity analysis strategy, organized step-by-step, was graphically depicted in Figure 5. The SPC-based two-way clustering was realized on a public server [26]. The corresponding programming codes for computing a function concept

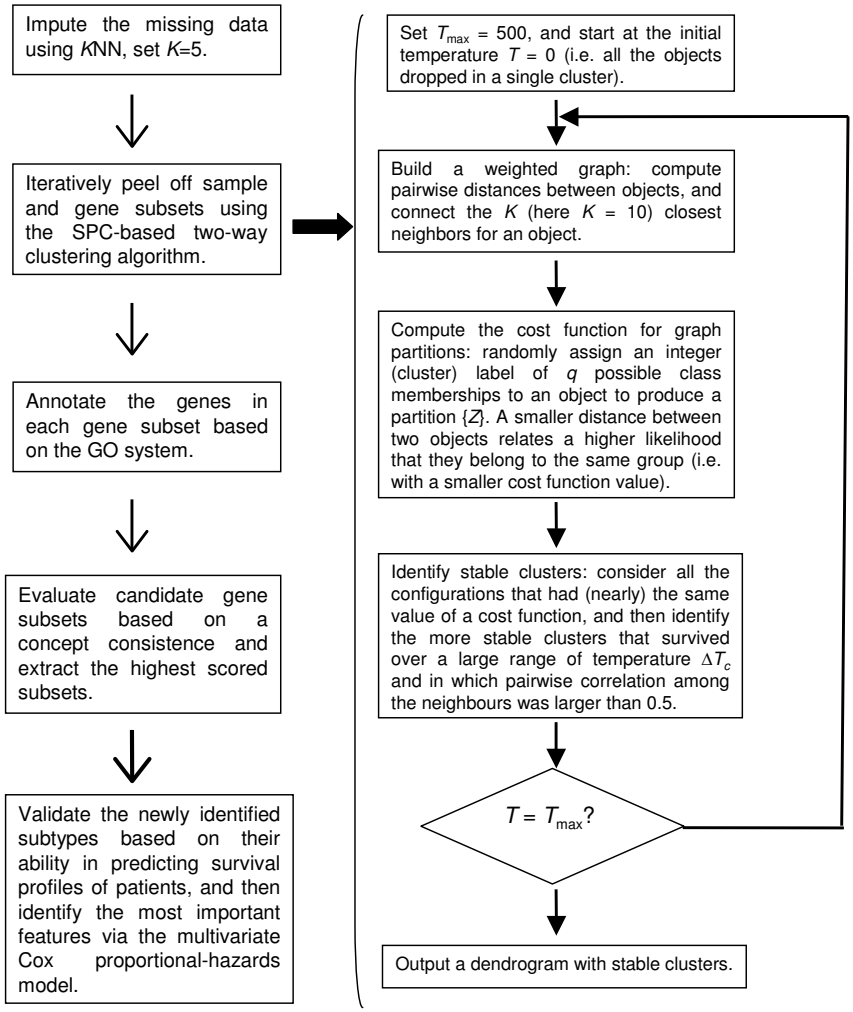

Figure 5
The graphic algorithm flow for the proposed SPC-based two-way clustering.

consistency score are available upon a written request to the authors. The hierarchical dendrogram resulted from the coupled two-way clustering was plotted by Treeview [27,28].

Abbreviations

diffuse large B-cell lymphoma (DLBCL), activated B-like DLBCL (AB-like DLBCL), germinal centre B-like DLBCL (GCB-like DLBCL), coupled two-way clustering (CTWC), super-paramagnetic clustering (SPC), GeneOntology (GO).

Competing interests

The author(s) declares that there are no competing interests.

Authors' contributions

This study was undertaken by a collaborative team of several institutes as indicated. WZ, LL, XL and SR conceived of the proposal of the study, conducted the study and drafted the manuscript. The remaining authors participated in writing the computing codes and applied the data mining strategy to the field datasets. All authors participated in reading, approving and revising the manuscript.

Additional material

Additional file 1

Table S1 – The functional annotations for the known genes included in G_2
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-332-S1.doc]

Additional file 2

Table S2 – The functional annotations for the known genes included in G_4
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-332-S2.doc]

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 60601010,30170515, 30370798, 30571034 and 30570424), the Heilongjiang Province Department of Education Outstanding Overseas Scientist grant (Grant No. 1055HG009), National Science Foundation of Heilongjiang Province (Grant Nos. ZJG0501, GB03C602-4 and F2004-02) and Health Department of Heilongjiang Province Key Project (2005-39).

References

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286(5439):**531-537.
2. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui CH, Evans WE, Naeve C, Wong L, Downing JR: **Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.** *Cancer Cell* 2002, **1(2):**133-143.
3. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406(6797):**747-752.
4. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL, Botstein D: **Repeated observation of breast tumor subtypes in independent gene expression data sets.** *Proc Natl Acad Sci U S A* 2003, **100(14):**8418-8423.
5. Sarwal M, Chua MS, Kambham N, Hsieh SC, Satterwhite T, Masek M, Salvatierra O Jr.: **Molecular heterogeneity in acute renal allograft rejection identified by DNA microarray profiling.** *N Engl J Med* 2003, **349(2):**125-138.
6. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.** *Proc Natl Acad Sci U S A* 2001, **98(24):**13790-13795.
7. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, Altman RB, Brown PO, Botstein D, Petersen I: **Diversity of gene expression in adenocarcinoma of the lung.** *Proc Natl Acad Sci U S A* 2001, **98(24):**13784-13789.
8. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, Ekman P, DeMarzo AM, Tibshirani R, Botstein D, Brown PO, Brooks JD, Pollack JR: **Gene expression profiling identifies clinically relevant subtypes of prostate cancer.** *Proc Natl Acad Sci U S A* 2004, **101(3):**811-816.
9. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr., Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403(6769):**503-511.
10. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltnane JM, Hurt EM, Zhao H, Averett L, Yang L, Wilson WH, Jaffe ES, Simon R, Klausner RD, Powell J, Duffey PL, Longo DL, Greiner TC, Weisenburger DD, Sanger WG, Dave BJ, Lynch JC, Vose J, Armitage JO, Montserrat E, Lopez-Guillermo A, Grogan TM, Miller TP, LeBlanc M, Ott G, Kvaloy S, Delabie J, Holte H, Krajci P, Stokke T, Staudt LM: **The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma.** *N Engl J Med* 2002, **346(25):**1937-1947.
11. Getz G, Domany E: **Coupled two-way clustering server.** *Bioinformatics* 2003, **19(9):**1153-1154.
12. Getz G, Gal H, Kela I, Notterman DA, Domany E: **Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data.** *Bioinformatics* 2003, **19(9):**1079-1089.
13. Getz G, Levine E, Domany E: **Coupled two-way clustering analysis of gene microarray data.** *Proc Natl Acad Sci U S A* 2000, **97(22):**12079-12084.
14. Tetko IV, Facius A, Ruepp A, Mewes HW: **Super paramagnetic clustering of protein sequences.** *BMC Bioinformatics* 2005, **6:**82.
15. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402(6761 Suppl):**C47-52.
16. Rives AW, Galitski T: **Modular organization of cellular networks.** *Proc Natl Acad Sci U S A* 2003, **100(3):**1128-1133.
17. Guo Z, Zhang T, Li X, Wang Q, Xu J, Yu H, Zhu J, Wang H, Wang C, Topol EJ, Rao S: **Towards precise classification of cancers based on robust gene functional expression profiles.** *BMC Bioinformatics* 2005, **6(1):**58.
18. 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(1-3):**25-33.
19. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification**

of biology. The Gene Ontology Consortium. *Nat Genet* 2000, **25(1):**25-29.
20. Lord PW, Stevens RD, Brass A, Goble CA: **Semantic similarity measures as tools for exploring the gene ontology.** *Pac Symp Biocomput* 2003:601-612.
21. Robinson PN, Wollstein A, Bohme U, Beattie B: **Ontologizing gene-expression microarray data: characterizing clusters with Gene Ontology.** *Bioinformatics* 2004, **20(6):**979-981.
22. Lymphoma/Leukemia Molecular Profiling Project: **Lymphoma/Leukemia Molecular Profiling Project.** [http://llmpp.nih.gov/lymphoma/index.shtml].
23. Wang D, Lv Y, Guo Z, Li X, Li Y, Zhu J, Yang D, Xu J, Wang C, Rao S, Yang B: **Effects of replacing the unreliable cDNA microarray measurements on the disease classification based on gene expression profiles and functional modules.** *Bioinformatics* 2006, **22(23):**2883-2889.
24. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR: **Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** *Nat Med* 2002, **8(1):**68-74.
25. The Broad Institute's Cancer Program Data Sets: **The Broad Institute's Cancer Program Data Sets.** [http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi].
26. The CTWC Server: **The Coupled Two Way Clustering algorithm (CTWC) Server .** [http://ctwc.weizmann.ac.il].
27. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.** *Proc Natl Acad Sci U S A* 1999, **96(12):**6745-6750.
28. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95(25):**14863-14868.
29. Li L, Jiang W, Li X, Moser KL, Guo Z, Du L, Wang Q, Topol EJ, Wang Q, Rao S: **A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset.** *Genomics* 2005, **85(1):**16-23.
30. Shen Y, Iqbal J, Xiao L, Lynch RC, Rosenwald A, Staudt LM, Sherman S, Dybkaer K, Zhou G, Eudy JD, Delabie J, McKeithan TW, Chan WC: **Distinct gene expression profiles in different B-cell compartments in human peripheral lymphoid organs.** *BMC Immunol* 2004, **5:**20.
31. Tedoldi S, Paterson JC, Cordell J, Tan SY, Jones M, Manek S, Dei Tos AP, Roberton H, Masir N, Natkunam Y, Pileri SA, Facchetti F, Hansmann ML, Mason DY, Marafioti T: **Jaw1/LRMP, a germinal centre-associated marker for the immunohistological study of B-cell lymphomas.** *J Pathol* 2006, **209(4):**454-463.
32. Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM: **A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma.** *Proc Natl Acad Sci U S A* 2003, **100(17):**9991-9996.
33. Lossos IS, Czerwinski DK, Alizadeh AA, Wechser MA, Tibshirani R, Botstein D, Levy R: **Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes.** *N Engl J Med* 2004, **350(18):**1828-1837.
34. Tagawa H, Suguro M, Tsuzuki S, Matsuo K, Karnan S, Ohshima K, Okamoto M, Morishima Y, Nakamura S, Seto M: **Comparison of genome profiles for identification of distinct subgroups of diffuse large B-cell lymphoma.** *Blood* 2005, **106(5):**1770-1777.
35. Behrens TW, Jagadeesh J, Scherle P, Kearns G, Yewdell J, Staudt LM: **Jaw1, A lymphoid-restricted membrane protein localized to the endoplasmic reticulum.** *J Immunol* 1994, **153(2):**682-690.
36. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB: **Large-scale analysis of the human and mouse transcriptomes.** *Proc Natl Acad Sci U S A* 2002, **99(7):**4465-4470.
37. Akasaka T, Lossos IS, Levy R: **BCL6 gene translocation in follicular lymphoma: a harbinger of eventual transformation to diffuse aggressive lymphoma.** *Blood* 2003, **102(4):**1443-1448.
38. Kopycka-Kedzierawski DT, Billings RJ: **A longitudinal study of caries onset in initially caries-free children and baseline salivary mutans streptococci levels: a Kaplan-Meier survival analysis.** *Community Dent Oral Epidemiol* 2004, **32(3):**201-209.
39. Cox DR: **Regression models and lifetables.** *JRStatSoc[B]* 1972, **34:**187-220.
40. Hans CP, Weisenburger DD, Greiner TC, Gascoyne RD, Delabie J, Ott G, Muller-Hermelink HK, Campo E, Braziel RM, Jaffe ES, Pan Z, Farinha P, Smith LM, Falini B, Banham AH, Rosenwald A, Staudt LM, Connors JM, Armitage JO, Chan WC: **Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray.** *Blood* 2004, **103(1):**275-282.
41. Gibbons FD, Roth FP: **Judging the quality of gene expression-based clustering methods using gene annotation.** *Genome Res* 2002, **12(10):**1574-1581.
42. Rhodes DR, Chinnaiyan AM: **Integrative analysis of the cancer transcriptome.** *Nat Genet* 2005, **37 Suppl:**S31-7.
43. Datta S, Datta S: **Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes.** *BMC Bioinformatics* 2006, **7:**397.
44. Hubert L, Arabie P: **Comparing partitions.** *Journal of Classification* 1985:193-218.
45. Ben-Hur A, Guyon I: **Detecting stable clusters using principal component analysis.** *Methods Mol Biol* 2003, **224:**159-182.
46. Bolshakova N, Azuaje F, Cunningham P: **An integrated tool for microarray data clustering and cluster validity assessment.** *Bioinformatics* 2005, **21(4):**451-455.
47. McLachlan GJ, Bean RW, Peel D: **A mixture model-based approach to the clustering of microarray expression data.** *Bioinformatics* 2002, **18(3):**413-422.
48. Blatt M, Wiseman S, Domany E: **Superparamagnetic clustering of data.** *Physical Review Letters* 1996, **76(18):**3251-3254.
49. Blockeel H, Bruynooghe M, Dzeroski S, Ramon J, Struyf J: **Hierarchical Multi-Classification.** In *Proceedings of the ACM SIGKDD 2002 Workshop on Multi-Relational Data Mining (MRDM 2002)* Edmonton, Canada ; 2002:21-35.
50. Altman DG: **Practical Statistics for Medical Research.** England , Chapman & Hall; 1991.