npg

Journal of Human Genetics (2012) 57, 642–653
© 2012 The Japan Society of Human Genetics All rights reserved 1434-5161/12 $32.00
www.nature.com/jhg

ORIGINAL ARTICLE

DBGSA: a novel method of distance-based gene set analysis

Jin Li[1,2,5], Limei Wang[3,5], Liangde Xu[1,5], Ruijie Zhang[1], Meilin Huang[1], Ke Wang[1], Jiankai Xu[1], Hongchao Lv[1], Zhenwei Shang[1], Mingming Zhang[1], Yongshuai Jiang[1], Maozu Guo[2,4] and Xia Li[1]

When compared with single gene functional analysis, gene set analysis (GSA) can extract more information from gene expression profiles. Currently, several gene set methods have been proposed, but most of the methods cannot detect gene sets with a large number of minor-effect genes. Here, we propose a novel distance-based gene set analysis method. The distance between two groups of genes with different phenotypes based on gene expression should be larger if a certain gene set is significantly associated with the given phenotype. We calculated the distance between two groups with different phenotypes, estimated the significant *P*-values using two permutation methods and performed multiple hypothesis testing adjustments. This method was performed on one simulated data set and three real data sets. After a comparison and literature verification, we determined that the gene resampling-based permutation method is more suitable for GSA, and the centroid statistical and average linkage statistical distance methods are efficient, especially in detecting gene sets containing more minor-effect genes. We believe that this distance-based method will assist us in finding functional gene sets that are significantly related to a complex trait. Additionally, we have prepared a simple and publically available Perl and R package (http://bioinfo.hrbmu.edu.cn/dbgsa or http://cran.r-project.org/web/packages/DBGSA/).

Journal of Human Genetics (2012) **57,** 642–653; doi:10.1038/jhg.2012.86; published online 12 July 2012

Keywords: average linkage statistical method; centroid statistical method; distance; gene set analysis; permutation

INTRODUCTION

With the development of biochip technology and microarrays that contain tens of thousands of genes, we can determine functional gene sets that are related to a phenotype using a gene function enrichment analysis method. There are two primary types of gene function enrichment analysis methods, individual gene analysis (IGA) and gene set analysis (GSA).[1–3] IGA identifies differentially expressed genes through a variety of methods and tests the difference of the proportion of differentially expressed genes between all genes and a given gene set.[4,5] GSA directly calculates gene subset scores using various statistical methods and calculates the significance level.[6] The IGA method requires an initial calculation of differentially expressed genes that is influenced by the statistical methods and their thresholds. Since the emergence of gene set enrichment analysis (GSEA), an increasing number of GSA approaches based on various statistical methods have been rapidly developed, such as GSEA,[7,8] globaltest,[9] SAM-GS,[10] GlobalANCOVA,[11] ADGO[12,13] and Bayesian network-based pathway analysis.[14]

Tian *et al.*[15] classified two types of null hypotheses that test whether a gene set displays a coordinated association with a phenotype of interest. The first type hypothesizes that the genes in a gene set have the same pattern of associations with the given phenotype when compared with the remaining genes (i.e., Q1). The second type hypothesizes that the gene set does not contain any genes that are associated with the given phenotype (i.e., Q2). Geoman and Buhlmann[16] termed competitive and self-contained methods based on Q1 and Q2, respectively. These methods have been widely used in previous studies. The genes can be divided into three categories, disease-related genes, minor-effect genes and disease-unrelated genes. The disease-related genes are significantly differentially expressed. The minor-effect genes individually exhibit marginal differential expression but may have a significant combined effect on the phenotypic outcome, such as disease.[17] The disease-unrelated genes exhibit no effect on the given phenotype. Usually, there are numerous genes that have a relatively minor-effect in complex diseases.[18,19] Therefore, it is critical to consider the minor-effect genes. Most of the gene function enrichment analysis methods can detect gene sets

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China; [2]School of Life Science and Technology, Harbin Institute of Technology, Harbin, China; [3]Center of Computer, Harbin Medical University, Harbin, China and [4]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[5]These authors contributed equally to this work.
Correspondence: Professor R Zhang or Professor Xia Li, College of Bioinformatics Science and Technology, Harbin Medical University, No 157. Road Baojian, Harbin 150081, China.
E-mail: zhangruijie2009@yahoo.com.cn (RZ) or lixia6@yahoo.com (XL)
or Professor M Guo, School of Life Science and Technology, Harbin Institute of Technology, Harbin, China.
E-mail: maozuguo@hit.edu.cn
Received 30 September 2011; revised 15 June 2012; accepted 19 June 2012; published online 12 July 2012

containing many disease-related genes well. However, they cannot detect the gene sets containing many minor-effect genes.

Here, we propose a novel distance-based gene set enrichment analysis method. We use the original gene expression data, not a summary statistic, in the analysis; therefore, this method utilizes each level of the gene expression data and is better able to detect functional sets, especially for gene sets containing more minor-effect genes. By focusing on gene expression, the distance between two groups with different phenotypes should be larger if a certain gene functional set is significantly associated with a particular phenotype. We use four point-to-point distance measures and two set-to-set distance measures to calculate the distance between two groups with different phenotypes, the case and control groups, by focusing on the gene expression profile of a given gene set. Next, we estimate a significant *P*-value for this gene set using permutation methods based on the two hypotheses above (Q1 and Q2) and perform multiple hypothesis testing adjustments using the false-discovery rate (FDR). We perform these methods on one simulated data set and three gene expression data sets and compare them with other methods.

In a parallel side study, we first transformed gene expression data to pathway activities using pathway-based microarray analysis methods, including the condition-responsive genes based (CORG-based),[20] negatively correlated feature sets with ideal markers (NCFS-i) and negatively correlated feature sets using the CORG-based method (NCFS-c) methods,[21,22] and then analyzed them with methods such as GSA or disease classification. In this manuscript, we present an improvement of these three methods for the detection of disease-related pathways.

MATERIALS AND METHODS

Data

Gene expression profiles. To analyze whether this method is feasible and effective, we used one simulated data set and three real data sets.

Simulated data set. We simulated a data set of 50 cases and 50 controls that included 10 000 genes for 100 samples. There was no exact proportion of the disease-related genes, the minor-effect genes and the disease-unrelated genes, but we thought there should be more minor-effect genes than significant disease-related genes. Therefore, we used proportions of 10% (1000 significant disease-related genes), 30% (3000 minor-effect genes) and 60% (6000 disease-unrelated genes) in this simulation. Reuben Thomas *et al.*[23] reported that a priori assumption of any of the considered univariate theoretical probability distributions across all probe sets was not valid. There was no assumption of probability distribution in our proposed methods. However, we needed to build a data set that included significantly differentially expressed genes and non-differentially expressed genes. The common and effective way to do this was to assume that the gene expression followed a normal distribution because we could easily control whether the genes were differentially expressed. The normal distribution was only used for the simulation purpose and we did not need any assumption of probability distribution in the proposed methods. We generated the disease-related genes as follows: 500 were downregulated genes that followed the normal distributions N (0,1) in the case group and N (1,1) in the control group. A total of 500 of the genes were upregulated and followed normal distributions N (1,1) in the case group and N (0,1) in the control group. All of the genes were significantly differentially expressed using a t-test with a significance level of 0.001. We generated the minor-effect genes as follows: 1500 of the genes followed normal distributions N (0,1) in the case group and N (0.5,1) in the control group. Fifteen hundred of the genes followed normal distributions N (0.5,1) in the case group and N (0,1) in the control group. About half of the genes were significantly differentially expressed using a t-test with a significance level of 0.001. So we considered them minor-effect genes. We generated the disease-unrelated genes as follows: 6000 genes followed a normal distribution N (0,1) in the case and control groups. None of the genes was significantly differentially expressed using a t-test with a significance level of 0.001.

Real data sets. Alzheimer's disease (AD) is a common neurodegenerative disease that severely affects the quality of life of the elderly. We selected an *AD* gene expression data set from Gene Expression Omnibus (GEO, GSE15222).[24] This data set included 24 350 probes, 363 samples, 187 controls and 176 cases. First, we performed pretreatments, including using the mean value of gene expression when multiple probes corresponded to one gene and removing missing data lines and outliers. We suggest performing these pretreatments before the GSA. Finally, we obtained expression values for 17 007 genes.

Non-small cell lung cancer (NSCLC) is a broad term for lung cancers that are not of the small-cell type. The three most common subtypes of NSCLC include adenocarcinoma (AC), squamous cell carcinoma (SCC) and large-cell carcinoma (LCC). We obtained two gene expression profiles of high-grade human NSCLC specimens. One data set, NSCLC I, included 58 samples (40 AC samples and 18 SCC samples, GSE10245).[25] After pretreating the data, we obtained expression values for 19 801 genes. The other data set, NSCLC II, included 28 samples (9 AC samples and 19 SCC samples, GSE27388).[26] After pretreating the data, we obtained expression values for 18 302 genes. We performed a Gene ontology (GO) functional set analysis for these data.

Gene sets. For the simulated data set, we constructed 1600 gene sets with different set sizes and proportions of different genes. We constructed 200 gene

Table 1 The presumed 200 gene sets with each set size

Gene sets description	*Number of gene sets*	*Proportion of disease-related genes (%)*	*Proportion of minor-effect genes (%)*	*Proportion of disease-unrelated genes (%)*
50 Presumed disease-related gene sets containing more disease-related genes	10	70	30[a]	
	10	60	40[a]	
	10	50	50[a]	
	10	40	60[a]	
	10	30	70[a]	
50 Presumed disease-related gene sets containing more minor-effect genes	10	10	90	0
	10	10	80	10
	10	10	70	20
	10	0	100	0
	10	0	90	10
100 presumed disease-unrelated gene sets	40	10	30	60
	30	0	30	70
	30	0	40	60

[a] This indicates the number is the sum of the proportions of minor-effect genes and disease-unrelated genes.

sets, 50 presumed disease-related gene sets containing more disease-related genes, 50 presumed disease-related gene sets containing more minor-effect and 100 presumed disease-unrelated gene sets, for set sizes of 10, 20, 30, 40, 50, 100, 150 and 200 genes. The detailed instructions are shown in Table 1.

For the real data sets, the gene sets were derived from the Molecular Signatures Database (MSigDB).[8] There are 6769 gene sets in MSigDB version 3.0 that are divided into five major collections. GO [27] is the most popular and widely used biomedical ontology. It is the de facto standard for effective functional annotation and enrichment analysis of high-throughput gene expression data sets. We used the GO gene sets, which are part of MSigDB v3.0 in this study. Because gene sets with too many or too few genes are uninformative, only 1401 GO gene sets with 10 to 500 genes were used.

Methods

Distance-based methods. First, we calculate the distance between the case and control groups by focusing on the gene expression for a given gene set. Next, we estimate the significant *P*-values for this gene set. And then we perform multiple hypothesis testing adjustments by FDR. The flow chart for this method is shown in Figure 1. The three key steps of the distance-based gene set analysis (DBGSA) method are described below.

Step 1: Calculate the distance between two groups with different phenotypes in a given gene set

First, we combine the gene expression profile and a given gene set from MSigDB to obtain a gene expression subset. Suppose that there are t individuals in the gene expression profile, which include t_1 individuals from the case group and t_2 individuals from the control group, and there are n genes in a gene set, which includes m genes in the gene expression profile. Here, we consider one person's gene expression values as a point in m dimensional space. Therefore, we can obtain t_1 and t_2 points in the case and control groups, respectively. The distance measures between the points and between the sets are defined below. We denote $d_{ij}0$ to be the distance between objects i and j and x_{ik} to be the kth gene expression value of person i.

The definitions of the distance between points are as follows.

Euclidean distance (-euc)

$$d_{\mathrm{ij}}(\mathrm{euc}) = \sqrt{\sum_{k=1}^{m} (x_{\mathrm{ik}} - x_{\mathrm{jk}})^2}$$

Statistical distance (-stat)

$$d_{\mathrm{ij}}(\mathrm{stat}) = \sqrt{\sum_{k=1}^{m} \frac{(x_{\mathrm{ik}} - x_{\mathrm{jk}})^2}{s_{\mathrm{kk}}}}$$

where s_{kk} is the sample variance of variable x_{ik}. The statistical distance is considered a weighted Euclidean distance using $k_1 = \frac{1}{s_{11}}, k_2 = \frac{1}{s_{22}}, \ldots\ k_m = \frac{1}{s_{\mathrm{mm}}}$ as the weight.

Manhattan distance (-man)

$$d_{\mathrm{ij}}(\mathrm{man}) = \sum_{k=1}^{m} |x_{\mathrm{ik}} - x_{\mathrm{jk}}|$$

Chebyshev distance (-max) $d_{\mathrm{ij}}(\mathrm{max}) = \max\limits_{1 \le k \le m} |x_{\mathrm{ik}} - x_{\mathrm{jk}}|$

The definition of the distance between the case and control sets is as follows.

Average linkage method (avelink-)

The average linkage method specifies that the distance between two sets is computed as the average distance between the objects from a set (case set G_{case}) and the objects from another set (control set G_{control}). The averaging is performed over all pairs (i, j) of objects, where i is an object from case set G_{case} and j is an object from control set $G_{control}$. This can be mathematically described as

$$D(\mathrm{Avelink}) = \frac{1}{t_1 t_2} \sum_{i \in G_{\mathrm{case}},\, j \in G_{\mathrm{control}}} d_{\mathrm{ij}}$$

where t_1 and t_2 are the sample numbers of set G_{case} and G_{control} and d_{ij} is the distance between i from G_{case} and j from G_{control}.

Centroid method (cent-)

The centroid method, which specifies the distance between two sets, is computed as the distance between the centroids of two sets. Mathematically, this method can be described as

$D(\mathrm{cent}) = d_{\overline{x_k}\ \overline{x_L}}$, where $\overline{x_k}$ and $\overline{x_L}$ are the centroid of sets G_{case} and G_{control}.

Each time, we select a definition of the distance measures between the points and between the sets. Therefore, we obtain eight different combinations of distance measures. We use the abbreviation of each combination of distance measures in the following analyses, such as avelink-euc for the average linkage Euclidean distance method. We denote the distance between the case and control sets as D^0.

Step 2: Estimate the significant level of the gene sets

We use permutation to estimate the significance level of the gene sets and perform two types of permutations according to Q1 and Q2.

Gene resampling-based permutation

There are two main methods in resampling theory, using subsets of available data (e.g., jackknifing) and drawing randomly with replacement from a set of data points (e.g., bootstrapping).[28–30] Because the presence of two or more of the same genes in one gene set is unlikely, we use the resampling method without replacement.

We randomly resample m_i genes from the gene expression profile and obtain a subset of gene expression profiles with s individuals and m_i genes. Next, we calculate the distances between the two groups with different phenotypes according to the method described in step 1. This procedure is repeated n_{per} times to obtain n_{per} distances randomly, which are denoted as $D^{1'}, D^{2'}, \ldots, D^{j'}, \ldots, D^{n'_{\mathrm{per}}}$.

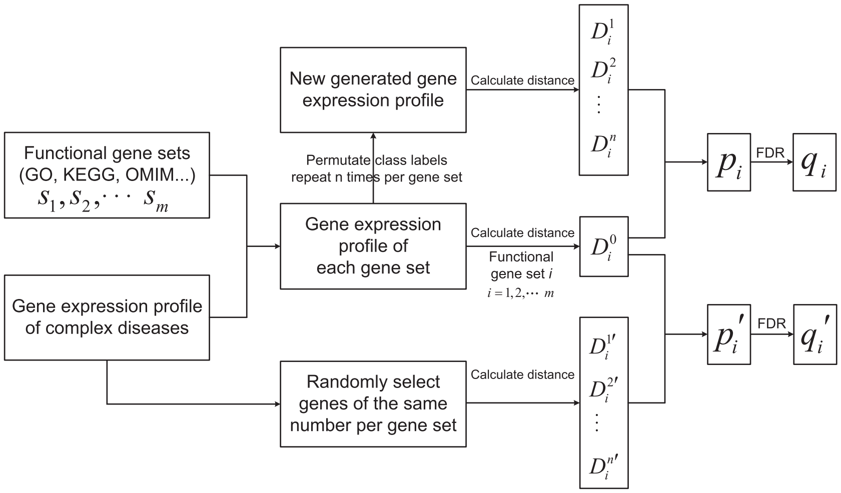

Figure 1 Flow chart of distance-based gene set analysis (DBGSA).

To screen out functional sets that are significantly associated with disease, we calculate *P*-values by a rank-based method as follows.

$$p' = 1 - \frac{\sum_{j=1}^{n_{\text{per}}} \delta'(j)}{n_{\text{per}}} \text{ where } \delta'(j) = \begin{cases} 0 & \text{if} \quad D^0 \leq D^{j'} \\ 1 & \text{if} \quad D^0 > D^{j'} \end{cases}$$

Label swapping-based permutation

For a specific gene expression profile set, we swap the label of each individual by controlling the same total number of cases and controls so as to obtain a new gene expression profile set. Next, we calculate the distance between the two groups with different phenotypes according to the method that was described in step 1. Similar to gene resampling, we repeat this procedure n_{per} times to randomly obtain n_{per} distances, and we denote them as $D^1, D^2, \ldots, D^j, \ldots, D^{n_{\text{per}}}$. We also calculate the *P*-value as follows:

$$p = 1 - \frac{\sum_{j=1}^{n_{\text{per}}} \delta(j)}{n_{\text{per}}} \text{ where } \delta(j) = \begin{cases} 0 & \text{if} \quad D^0 \leqslant D^j \\ 1 & \text{if} \quad D^0 > D^j \end{cases}$$

Step 3: Adjusted for multiple hypothesis testing

The estimated significance level should be adjusted to account for multiple hypothesis testing when thousands of gene sets are tested. The calculation of FDR has been shown to be an effective method. We use fdrtool[31] to estimate the tail area-based FDR (Fdr) and density-based local FDR in this study.

Using these three steps, we can determine whether a gene set is significantly related to a trait.

Gene set enrichment analysis. GSEA is a widely used gene set analysis method. We use a gene resampling-based GSEA method for comparison in this study.

CORG-based method, NCFS-i method and NCFS-c method. CORG-based, NCFS-i and NCFS-c methods are efficient pathway-based microarray analysis methods.[20–22] We present an improvement based on these three methods for the detection of disease-related pathways. First, we use these original methods to select a subset for each gene set. Next, we use five-fold cross validation to calculate the accuracy of the subset in disease classification. Finally, we classify the top gene sets with high accuracy as the disease-related gene sets. Instead of *P*-values, we choose the significance threshold to be a classification of accuracy.

Precision, recall and the F-measure (F_1). We use precision, recall and the *F*-measure (F_1) to evaluate different methods for detecting disease-related gene sets. These terms are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F_1 = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

The terms TP, TN, FP and FN represent true positives, true negatives, false positives and false negatives, respectively.

Overlap coefficient. While measuring the similarity of two sets with large differences in set sizes, we should pay more attention to the smaller set. The overlap coefficient[32,33] is a proper similarity measure in this situation that computes the overlap between two sets and is defined as follows:

$$\text{overlap}(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$$

where *X* and *Y* are two sets and |*X*| indicates the set size of *X*.

RESULTS AND DISCUSSION

In this section, the proposed method was compared with other methods. First, we evaluated the label swapping-based permutation method and found that it was improper. Next, we evaluated the gene resampling-based permutation method in the simulated data set and found that the avelink-stat and cent-stat methods were appropriate for GSA. Additionally, we compared these methods with GSEA, the CORG-based method, the NCFS-i method and the NCFS-c method using various evaluation measures in the simulated data set and the real data sets. The details are shown in Figure 2.

Results from the label swapping-based permutation

For the AD data, we found that most of the functional sets were significantly associated with AD by our label swapping-based permutation method; the results are shown in Table 2.

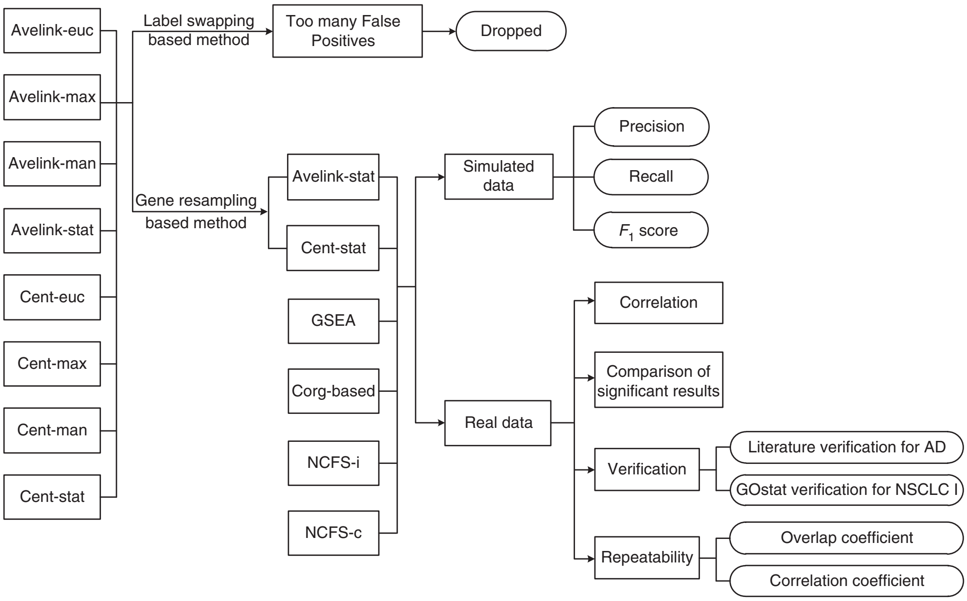

Figure 2 Flow chart of the comparison and evaluation.

Table 2 Result of the number of significant sets from the label swapping-based permutation

Method	*Number of significant gene sets,* P*<0.001 (%)*	*Method*	*Number of significant gene sets,* P*<0.001 (%)*
Avelink-euc	1307 (93.3)[a]	Cent-euc	1197 (85.4)[a]
Avelink-man	1373 (98.0)[a]	Cent-man	1322 (94.4)[a]
Avelink-max	1241 (88.6)[a]	Cent-max	1053 (75.2)[a]
Avelink-stat	1400 (99.9)[a]	Cent-stat	1392 (99.4)[a]

[a] The number in the parentheses indicates the proportion of detected significant gene sets in all the gene sets.

Table 3 The number of significant results by selecting different numbers of disease-unrelated and -related genes in the simulated data

Number of disease-unrelated and -related genes	*Number of significant results at* P*<0.001*
9–1	10
8–2	10
19–1	10
18–2	10
49–1	7
48–2	10
99–1	7
98–2	10
199–1	2
198–2	9

Table 4 The number of significant results by selecting different numbers of genes from the AD data

Number of genes selected from the AD data	*Number of significant results at* P*<0.001*
10	85
20	89
50	88
100	99
200	100

Abbreviation: AD, Alzheimer's disease.

To understand these results, we performed a simulation experiment. We selected different numbers of disease-unrelated genes and disease-related genes from the simulated data set to form gene sets (gene set sizes ranging from 10 to 200 genes in 5 gradients). And then we calculated *P*-values using the avelink-euc method by performing 1000 permutations. We repeated this procedure 10 times for each case and observed the number of significant results. These results are shown in Table 3.

If there are more disease-related genes in a gene set with a given set size, then there is a higher probability that the gene set detected is disease-related. In the simulated data, the proportion of the disease-related genes is 10%. However, the results indicate that although there were only two disease-related genes in a tested gene set containing 200 genes (the proportion of the disease-related genes is just 1%), we could still detect the gene set significantly related to disease.

We also evaluated this method in the real AD data set. We randomly selected several genes (gene set sizes ranging from 10 to 200 genes in 5 gradients) to form a putative functional category, calculated *P*-values using the avelink-euc distance method by performing 1000 permutations, and repeated the procedure 100 times for each case. Next, we calculated the number of significant results using a significance level of 0.001. These results are shown in Table 4.

Based on the results, even the random selection of genes from the AD data set would produce significant results with a probability of >85%, indicating that many false disease-related gene sets (false positives) may be found. The simulated and real data indicated that this label swapping-based permutation method was too sensitive and may result in high false-positive rates in the detection of significant disease-related functional gene sets.

Results from the gene resampling-based permutation

Comparison of the simulated data set. We performed gene resampling-based GSEA, the CORG-based method, the NCFS-i method, the NCFS-c method and our proposed gene resampling-based permutation method on the simulated data set. We selected $P<0.05$ as the significance threshold for the gene resampling-based GSEA and our gene resampling-based permutation method. Because there is no empirical threshold for the CORG-based, NCFS-i and NCFS-c methods, we selected a 5-fold accuracy >0.9, >0.85 and >0.8 as thresholds. We counted the number of statistically significant disease-related gene sets using these methods and calculated the precision, recall and F_1 scores. These results are shown in Table 5.

We aimed to select the most efficient of our eight proposed methods. From the results, we found that the avelink-stat method performed best out of the average linkage methods, and the cent-euc, cent-man and cent-stat methods performed similarly out of the centroid methods. Because the variance in our simulated data set was set to 1 for all the genes, we replaced 10 of the minor-effect genes with larger variances to test the robustness of our methods for unnormalized data. In these 10 genes, the expression followed a normal distribution N (10,1000) in the case group and a normal distribution N (0,1000) in the control group. We constructed 100 disease-unrelated gene sets within 100 genes as follows. In each gene set, we randomly selected 10 disease-related genes (10%), 1 minor-effect gene with a variance of 1000 and 29 minor-effect genes with a variance of 1 (30%), and 60 disease-unrelated genes (60%). We suggested that these 100 gene sets were unrelated to disease because the proportions of different gene types were the same as the total data set. We performed the proposed gene resampling-based methods on this data set, selected $P<0.05$ as the significance threshold and counted the number of significant disease-related gene sets. The results indicate that only the avelink-stat and cent-stat methods are robust for the dimensions of the gene expression values; therefore, we used these two methods in the following analyses. The results are shown in Table 5.

Next, we compared the avelink-stat, cent-stat, CORG-based, NCFS-i and NCFS-c methods and gene resampling-based GSEA. The gene set size influenced the results, such that more significant results were obtained with larger gene set sizes. In the CORG-based, NCFS-i and NCFS-c methods, the influences of the threshold in the gene sets with different gene set sizes differed. When the gene set size was small, more truly significant disease-related gene sets were obtained using a lower acc threshold (acc>0.8). But when the gene set size was large, too many false significant disease-related gene sets were obtained using a lower acc threshold (97, 99 and 99 false

Table 5 Results in the simulated data set using different methods

		Avelin	*Avelin*	*Avelin*	*Avelin*														
Set size	*Results*	*k-euc*	*k-max*	*k-man*	*k-stat*	*Cent-euc*	*Cent-max*	*Cent-man*	*Cent-stat*	*GSEA*	*CORG-based*	*NCFS-i*	*NCFS-c*	*CORG-based*	*NCFS-i*	*NCFS-c*	*CORG-based*	*NCFS-i*	*NCFS-c*
		$P<0.05$	$P<0.05$	$P<0.05$	$P<0.05$	$P<0.05$	$P<0.05$	$P<0.05$	$P<0.05$	$P<0.05$	*acc>0.9*	*acc>0.9*	*acc>0.9*	*acc>0.85*	*acc>0.85*	*acc>0.85*	*acc>0.8*	*acc>0.8*	*acc>0.8*
10	Positive 1	39	35	39	44	42	13	40	43	10	2	10	7	12	27	25	33	43	42
	Positive 2	5	7	5	7	8	2	23	11	0	0	0	1	1	1	2	3	16	17
	Positive 3	1	1	1	0	0	1	0	0	7	0	0	0	0	0	0	2	5	3
	Precision	0.98	0.98	0.98	1.00	1.00	0.94	1.00	1.00	0.59	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.92	0.95
	Recall	0.44	0.42	0.44	0.51	0.50	0.15	0.63	0.54	0.10	0.02	0.10	0.08	0.13	0.28	0.27	0.36	0.59	0.59
	F_1 score	0.61	0.59	0.61	0.68	0.67	0.26	0.77	0.70	0.17	0.04	0.18	0.15	0.23	0.44	0.43	0.52	0.72	0.73
20	Positive 1	49	43	48	48	49	12	47	48	7	13	25	25	40	46	45	48	50	50
	Positive 2	11	9	14	18	13	1	43	20	2	0	0	0	5	16	10	16	39	34
	Positive 3	1	1	1	0	0	1	0	0	5	0	0	0	0	3	5	14	18	19
	Precision	0.98	0.98	0.98	1.00	1.00	0.93	1.00	1.00	0.64	1.00	1.00	1.00	1.00	0.95	0.92	0.82	0.83	0.82
	Recall	0.60	0.52	0.62	0.66	0.62	0.13	0.90	0.68	0.09	0.13	0.25	0.25	0.45	0.62	0.55	0.64	0.89	0.84
	F_1 score	0.75	0.68	0.76	0.80	0.77	0.23	0.95	0.81	0.16	0.23	0.40	0.40	0.62	0.75	0.69	0.72	0.86	0.83
30	Positive 1	45	46	45	48	49	10	49	48	17	22	36	38	48	49	47	50	50	50
	Positive 2	8	6	12	27	19	0	49	31	3	0	3	2	11	23	14	31	45	42
	Positive 3	0	0	0	0	0	5	0	0	4	0	2	1	4	21	16	23	48	41
	Precision	1.00	1.00	1.00	1.00	1.00	0.67	1.00	1.00	0.83	1.00	0.95	0.98	0.94	0.77	0.79	0.78	0.66	0.69
	Recall	0.53	0.52	0.57	0.75	0.68	0.10	0.98	0.79	0.20	0.22	0.39	0.40	0.59	0.72	0.61	0.81	0.95	0.92
	F_1 score	0.69	0.68	0.73	0.86	0.81	0.17	0.99	0.88	0.32	0.36	0.55	0.57	0.72	0.75	0.69	0.79	0.78	0.79
40	Positive 1	48	46	48	50	50	12	50	50	31	26	48	44	48	50	50	50	50	50
	Positive 2	22	11	23	33	24	1	49	36	5	0	8	8	14	25	22	39	44	43
	Positive 3	3	1	3	0	0	2	0	0	1	1	4	1	12	29	23	35	58	54
	Precision	0.96	0.98	0.96	1.00	1.00	0.87	1.00	1.00	0.97	0.96	0.93	0.98	0.84	0.72	0.76	0.72	0.62	0.63
	Recall	0.70	0.57	0.71	0.83	0.74	0.13	0.99	0.86	0.36	0.26	0.56	0.52	0.62	0.75	0.72	0.89	0.94	0.93
	F_1 score	0.81	0.72	0.82	0.91	0.85	0.23	0.99	0.92	0.53	0.41	0.70	0.68	0.71	0.74	0.74	0.79	0.75	0.75
50	Positive 1	49	46	49	50	50	7	50	50	30	34	46	46	50	50	50	50	50	50
	Positive 2	26	14	25	42	33	0	49	43	9	4	10	18	23	33	30	45	47	47
	Positive 3	1	1	1	0	0	2	0	0	1	0	5	6	17	36	25	42	65	61
	Precision	0.99	0.98	0.99	1.00	1.00	0.78	1.00	1.00	0.98	1.00	0.92	0.91	0.81	0.70	0.76	0.69	0.60	0.61
	Recall	0.75	0.60	0.74	0.92	0.83	0.07	0.99	0.93	0.39	0.38	0.56	0.64	0.73	0.83	0.80	0.95	0.97	0.97
	F_1 score	0.85	0.75	0.85	0.96	0.91	0.13	0.99	0.96	0.56	0.55	0.70	0.75	0.77	0.76	0.78	0.80	0.74	0.75
100	Positive 1	50	48	50	50	50	9	50	50	43	49	50	50	50	50	50	50	50	50
	Positive 2	33	21	35	46	41	1	50	49	14	8	35	31	36	44	47	47	50	50
	Positive 3	0	2	0	0	0	2	0	0	2	9	30	29	39	61	52	76	87	86
	Precision	1.00	0.97	1.00	1.00	1.00	0.83	1.00	1.00	0.97	0.86	0.74	0.74	0.69	0.61	0.65	0.56	0.53	0.54
	Recall	0.83	0.69	0.85	0.96	0.91	0.10	1.00	0.99	0.57	0.57	0.85	0.81	0.86	0.94	0.97	0.97	1.00	1.00
	F_1 score	0.91	0.81	0.92	0.98	0.95	0.18	1.00	0.99	0.72	0.69	0.79	0.77	0.76	0.74	0.78	0.71	0.70	0.70
150	Positive 1	50	49	50	50	50	12	50	50	44	50	50	50	50	50	50	50	50	50
	Positive 2	41	25	44	50	46	3	50	50	23	19	32	38	43	50	49	50	50	49
	Positive 3	1	1	1	0	0	2	0	0	5	16	48	50	53	79	78	90	99	95
	Precision	0.99	0.99	0.99	1.00	1.00	0.88	1.00	1.00	0.93	0.81	0.63	0.64	0.64	0.56	0.56	0.53	0.50	0.51
	Recall	0.91	0.74	0.94	1.00	0.96	0.15	1.00	1.00	0.67	0.69	0.82	0.88	0.93	1.00	0.99	1.00	1.00	0.99
	F_1 score	0.95	0.85	0.96	1.00	0.98	0.26	1.00	1.00	0.78	0.75	0.71	0.74	0.76	0.72	0.71	0.69	0.67	0.67

Table 5 (Continued)

Set size	Results	Avelin k-euc	Avelin k-max	Avelin k-man	Avelin k-stat	Cent-euc	Cent-max	Cent-man	Cent-stat	GSEA	CORG-based	NCFS-i	NCFS-c	CORG-based	NCFS-i	NCFS-c	CORG-based	NCFS-i	NCFS-c
		P<0.05	P<0.05	P<0.05	P<0.05	P<0.05	P<0.05	P<0.05	P<0.05	P<0.05	acc>0.9	acc>0.9	acc>0.9	acc>0.85	acc>0.85	acc>0.85	acc>0.8	acc>0.8	acc>0.8
200	Positive 1	50	50	50	50	50	10	50	50	49	49	50	50	50	50	50	50	50	50
	Positive 2	45	24	45	50	46	1	50	50	28	23	45	42	45	50	50	50	50	50
	Positive 3	0	0	0	0	0	1	0	0	0	30	48	50	60	85	81	97	99	99
	Precision	1.00	1.00	1.00	1.00	1.00	0.92	1.00	1.00	1.00	0.71	0.66	0.65	0.61	0.54	0.55	0.51	0.50	0.50
	Recall	0.95	0.74	0.95	1.00	0.96	0.11	1.00	1.00	0.77	0.72	0.95	0.92	0.95	1.00	1.00	1.00	1.00	1.00
	F_1 score	0.97	0.85	0.97	1.00	0.98	0.20	1.00	1.00	0.87	0.71	0.78	0.76	0.75	0.70	0.71	0.67	0.67	0.67
	Positive 4	39	49	49	0	36	53	46	0	3	22	64	67	87	98	95	99	100	99
	Specificity	0.61	0.51	0.51	1.00	0.64	0.47	0.54	1.00	0.97	0.78	0.36	0.33	0.13	0.02	0.05	0.01	0.00	0.01

Abbreviations: Avelink, average linkage; Cent, centroid; euc, Euclidean distance; GSEA, gene set enrichment analysis; stat, statistical distance.
Positive 1: the number of significant disease-related gene sets at the given significance level in the 50 gene sets containing more disease-related genes.
Positive 2: the number of significant disease-related gene sets at the given significance level in the 50 gene sets containing more minor-effect genes.
Positive 3: the number of significant disease-related gene sets at the given significance level in the 100 disease-unrelated gene sets.
Positive 4: the number of significant disease-related gene sets at the given significance level in the 100 disease-unrelated gene sets with unnormalized data.
While calculating Precision, Recall and F_1 score, we classified the results into two (disease-related, disease-unrelated) categories.
TP = Positive 1 + Positive 2.
FP = Positive 3.
TN = 100 − FP = 100-Positive 3.
FN = 100 − TP = 100-Positive 1-Positive 2.
The specificity was defined as the true negatives (TN) divided by the total negatives (TN + FP), and it was calculated as (100 −positive4)/100 here.

positives in 100 negative gene sets when the gene set size is 200 using acc >0.8 as the threshold). In the gene set analysis, we wanted more accurate results with a lower FDR. Therefore, we selected the higher acc threshold (acc>0.9) for these three methods. For precision, the avelink-stat and cent-stat methods resulted in higher values than GSEA when the gene set sizes were small (<50); the avelink-stat and cent-stat methods resulted in higher values than the CORG-based, NCFS-i and NCFS-c methods when the gene set sizes were large (>100) because the CORG-based, NCFS-i and NCFS-c methods resulted in more false-negative gene sets than the proposed two methods (e.g., 30, 48 and 50 false negatives in these three methods vs no false negatives in our two methods when the gene set size is 200). The overall precisions (1600 gene sets) of these six methods were 1, 1, 0.86, 0.92, 0.85 and 0.86, respectively. For recall, the avelink-stat and cent-stat methods resulted in higher values than the four other methods for all gene set sizes. The overall recall values for these six methods were 0.83, 0.85, 0.39, 0.37, 0.56 and 0.56, respectively, because our methods resulted in more true-positive gene sets, especially those gene sets containing more minor-effect genes (e.g., 42 and 43 true positive gene sets containing more minor-effect genes in our methods vs 9, 4, 10 and 18 in the other 4 methods when the gene set size was 50). The overall F_1 scores of these six methods were 0.91, 0.92, 0.54, 0.53, 0.68 and 0.68, respectively, which demonstrates that our methods were best. We also compared these methods in negative gene sets with unnormalized data. The significant false-positive gene sets from the 100 negative gene sets were 0, 0, 3, 22, 64 and 67 using these six methods, respectively. Therefore, our two methods and gene resampling-based GSEA were robust for the unnormalized data. Taking into consideration the overall precision, recall and F_1 score of the 1600 gene sets, we believe that both the avelink-stat method (1, 0.83 and 0.91, respectively) and the cent-stat method (1, 0.85 and 0.92, respectively) are better suited for GSA and perform better than other methods, especially in detecting disease-related gene sets containing more minor-effect genes and in their robustness for handling unnormalized data. We obtained a significantly higher overall recall of the 1600 gene sets using the cent-stat method (0.85) compared with the avelink-stat method (0.83) or other methods (GSEA, 0.39; the CORG-based, 0.37; NCFS-i, 0.56; and NCFS-c, 0.5625). Therefore, the cent-stat method was more sensitive than the avelink-stat and other methods. Thus, we performed these two methods on three real data sets.

A comparison of the avelink-stat method and the cent-stat method in real data sets

We calculated the average distance between each point from two gene sets for the average linkage distance method. Next, we determined the centroid of each gene set and calculated the distances between the two centroids for the centroid distance method. To further compare the detection ability of the two methods, we selected a Fdr<0.05 as the significance threshold of association between the gene sets and traits in the real data sets. For the AD data, we obtained 10 statistically significant functional sets using the avelink-stat method and 116 sets using the cent-stat method; 8 of these sets were included in both of the methods. For the NSCLC I data, we obtained 91 statistically significant functional sets using the avelink-stat method and 129 sets using the cent-stat method; 68 of these sets were included in both of the methods. For the NSCLC II data, we obtained 321 statistically significant functional sets using the avelink-stat method and 495 sets using the cent-stat method; 304 of these sets were included in both of the methods. In these examples, we obtained more significant

Table 6 Number of significant functional gene sets in the real data sets using different methods

		Verified by			*Verified by*				*Overlap*
Methods	*AD*	*literature*	*Precision*	*NSCLC I*	*GOstat*	*Precision*	*Recall*	*NSCLC II*	*coefficient (%)*
Avelink-stat	10	9	0.90	91	10	0.11	0.43	321	52.8
Cent-stat	116	68	0.59	129	11	0.09	0.48	495	70.5
Avelink-stat and Cent-stat	8	8	1.00	68	10	0.15	0.43	304	85.3
GSEA	42	22	0.52	63	7	0.11	0.30	399	77.8
Avelink-stat and GSEA	1	1	1.00	31	6	0.19	0.26	189	51.6
Cent-stat and GSEA	23	13	0.57	51	7	0.14	0.30	278	60.8
Avelink-stat, Cent-stat and GSEA	1	1	1.00	30	6	0.20	0.26	189	50.0
CORG-based method	0	0	—	194	5	0.03	0.22	393	64.4
NCFS-i	0	0	—	181	8	0.04	0.35	438	63.0
NCFS-c	0	0	—	179	9	0.05	0.39	425	63.1
CORG-based method, NCFS-i and NCFS-c	0	0	—	63	2	0.03	0.09	268	60.3
Avelink-stat and CORG-based method	0	0	—	39	5	0.13	0.22	192	51.3
Avelink-stat and NCFS-i	0	0	—	30	5	0.17	0.22	202	53.3
Avelink-stat and NCFS-c	0	0	—	31	7	0.23	0.30	205	58.1
Avelink-stat, CORG-based method, NCFS-i and NCFS-c	0	0	—	12	2	0.17	0.09	152	58.3
Cent-stat and CORG-based method	0	0	—	45	5	0.11	0.22	253	66.7
Cent-stat and NCFS-i	0	0	—	34	5	0.15	0.22	270	67.6
Cent-stat and NCFS-c	0	0	—	34	7	0.21	0.30	275	73.5
Cent-stat, CORG-based method, NCFS-i and NCFS-c	0	0	—	17	2	0.12	0.09	197	70.6
Avelink-stat, Cent-stat, CORG-based method, NCFS-i and NCFS-c	0	0	—	9	2	0.22	0.09	151	55.6

Abbreviations: AD, Alzheimer's disease; Avelink, average linkage; Cent, centroid; euc, Euclidean distance; GSEA, gene set enrichment analysis; NSCLC, Non-small cell lung cancer; stat, statistical distance.
Precision was calculated as the total number of actual disease-related gene sets found by the method (verified by literature or GOstat), divided by total disease-related gene sets found by the method.
Recall was calculated as the total number of actual disease-related gene sets found by the method (verified by GOstat), divided by total disease-related gene sets found by GOstat.

functional sets using the cent-stat method than the avelink-stat method. Detailed results are shown in Table 6.

A comparison of the distance-based method and other methods in real data sets

In addition, we performed gene resampling-based GSEA, the CORG-based method, the NCFS-i method and the NCFS-c method using these three data sets. We calculated the correlation coefficients of the GSEA, avelink-stat and cent-stat methods. The results are shown in Table 7. The correlation between the cent-stat method and the avelink-stat method was greater than the correlations between the cent-stat method and GSEA and between the avelink-stat method and GSEA. The detailed computational results for the AD data are shown in Supplementary Table 1.

We selected a Fdr <0.05 as the significance threshold for the GSEA analysis. Moreover, we selected an acc>0.9 as the significance threshold for the CORG-based, NCFS-i and NCFS-c methods based on the analysis of the threshold choice in the simulated data.

For the AD data, we obtained 42 statistically significant functional gene sets using GSEA. One of these sets was shared between GSEA and the avelink-stat method, and 23 sets were shared between GSEA and the cent-stat method; one set was shared by all three methods. The results are shown in Table 6. More functional sets were obtained using the cent-stat method than GSEA, which included more detailed portions of the functional sets. For example, we obtained 'regulation of transcription, DNA dependent' using GSEA. We obtained some additional depth nodes by the cent-stat method, such as 'regulation of transcription factor activity' and 'negative regulation of transcription DNA dependent'. We did not obtain any statistically significant functional gene sets using the CORG-based, NCFS-i and NCFS-c methods. Even when we set the acc >0.8 as the significance threshold, we could not obtain any statistically significant functional gene set using these methods.

For the NSCLC I data, we obtained 63 statistically significant functional sets using GSEA. A total of 31 of these sets were shared between GSEA and the avelink-stat method, 51 were shared between GSEA and the cent-stat method, and 30 were shared by all three methods. For the NSCLC II data, we obtained 399 statistically significant functional sets using GSEA. Overall, 189 of these sets were shared between GSEA and the avelink-stat method, 278 were shared between GSEA and the cent-stat method, and 189 were shared by all three methods. These results are shown in Table 6 and show that more significant disease-related gene sets are obtained using the cent-stat method than gene resampling-based GSEA.

For the NSCLC I data, we obtained 194, 181 and 179 statistically significant functional sets using the CORG-based, NCFS-i and NCFS-c methods, respectively. A total of 63 of these sets were shared by the three methods. For the NSCLC II data, we obtained 393, 438 and 425 statistically significant functional sets using the CORG-based, NCFS-i and NCFS-c methods, respectively. Overall, 268 of these sets were shared by the three methods. These results are shown in Table 6. We obtained as many significant gene sets using the CORG-based, NCFS-i and NCFS-c methods as using the distance-based methods.

Literature verification

We performed a literature verification to determine whether the significant functional gene sets obtained from different methods were associated with the trait. For the AD data, we acquired 10 significant functional sets from the avelink-stat method. We found that 9 out of the 10 gene sets were already recognized in a large number of AD

Table 7 Correlation of *P*-values (or acc values) among the methods in the real data sets

	GSEA	*Avelink-stat*	*Cent-stat*
(a) Correlation of P*-values in the AD data*			
GSEA	1.0000	0.2729	0.4979
Avelink-stat	0.2612	1.0000	0.6801
Cent-stat	0.4735	0.6532	1.0000

	GSEA in I	*Avelink-stat in I*	*Cent-stat in I*	*GSEAin II*	*Avelink-stat in II*	*Cent-stat in II*
(b) Correlation of P*-values in the NSCLC data I and NSCLC data II*						
GSEA in I	1.0000	0.3104	0.3664	0.1326	0.0607	0.0704
Avelink-stat in I	0.2175	1.0000	0.8547	0.4422	0.4421	0.4754
Cent-stat in I	0.2781	0.8297	1.0000	0.3737	0.4513	0.4785
GSEA in II	0.0485	0.3822	0.2772	1.0000	0.4723	0.5726
Avelink-stat in II	0.0239	0.4290	0.4430	0.3140	1.0000	0.9058
Cent-stat in II	0.0298	0.4192	0.4365	0.3701	0.8595	1.0000

	CORG in I	*NCFS-i in I*	*NCFS-c in I*	*CORG in II*	*NCFS-i in II*	*NCFS-c in II*
(c) Correlation of acc values in the NSCLC data I and NSCLC data II						
CORG in I	1.0000	0.7280	0.6900	0.4015	0.4354	0.4126
NCFS-i in I	0.7378	1.0000	0.8667	0.4027	0.3957	0.3910
NCFS-c in I	0.6954	0.8736	1.0000	0.3638	0.3713	0.3731
CORG in II	0.4316	0.4230	0.3891	1.0000	0.7683	0.7244
NCFS-i in II	0.4483	0.4028	0.3818	0.7813	1.0000	0.8760
NCFS-c in II	0.4258	0.3982	0.3844	0.7374	0.8818	1.0000

Abbreviations: AD, Alzheimer's disease; Avelink, average linkage; Cent, centroid; GSEA, gene set enrichment analysis; NSCLC, non-small cell lung cancer; stat, statistical distance.
In these tables, Pearson's correlation coefficients were shown in the lower triangular table, and Spearman correlation coefficients were shown in the upper triangular table.

Table 8 Literature verification of the top 10 significant gene sets that were obtained from different methods in AD

Avelink-stat method	*GSEA*	*CORG-based method*	*NCFS-i*	*NCFS-c*
Mitochondrion[34][a]	**Mitochondrion**[34][a]	**Biosynthetic process**[35][a]	**Biosynthetic process**[35][a]	**Biosynthetic process**[35][a]
Energy derivation by oxidation of organic compounds[36,37][a]	Aerobic respiration[38][a]	Cytoskeleton[39][a]	**Muscle development**[40][a]	**Muscle development**[40][a]
Endoplasmic reticulum part[36,41][a]	Cellular respiration[42][a]	RNA processing[43][a]	**Endoplasmic reticulum**[44,45][a]	**Endoplasmic reticulum**[44,45][a]
Organelle membrane[46][a]	Mitochondrial inner membrane[47][a]	Serine hydrolase activity[48][a]	**Membrane lipid metabolic process**[49][a]	**Membrane lipid metabolic process**[49][a]
Negative regulation of programmed cell death[50][a]	Mitochondrial membrane part[51][a]	Serine-type endopeptidase activity[52][a]	**Nervous system development**[53][a]	**Nervous system development**[53][a]
Membrane organization and biogenesis[54][a]	Mitochondrial membrane[51][a]	Macromolecule biosynthetic process[35][a]	Phospholipid metabolic process[55][a]	UDP glycosyltransferase activity[56][a]
Regulation of binding[46][a]	Mitochondrial part[57][a]	Cytoskeleton organization and biogenesis[58][a]	**Transferase activity transferring glycosyl groups**[59][a]	**Transferase activity transferring glycosyl groups**[59][a]
Regulation of molecular function[60][a]	Organelle inner membrane[61][a]	RNA binding	Cellular biosynthetic process[35][a]	Positive regulation of developmental process
Regulation of DNA binding[62][a]	Mitochondrial envelope	Transport vesicle	**Electron carrier activity**	**Electron carrier activity**
DNA catabolic process	Proton transporting two sector ATPase complex	Serine-type peptidase activity	Protein heterodimerization activity	Substrate-specific transmembrane transporter activity

Abbreviations: AD, Alzheimer's disease; Avelink, average linkage; GSEA, gene set enrichment analysis; stat, statistical distance.
The overlapping gene sets of two methods were shown in bold.
[a]These functional sets were verified in the literature.

literature sources. Additionally, we searched for the top 10 significant GO terms obtained by GSEA, CORG-based method, NCFS-i and NCFS-c, respectively.[34–62] We confirmed that eight of them from GSEA, seven of them from CORG-based method, eight of them from NCFS-i and seven of them from NCFS-c were verified in the literature. There was a linear relation between the recall and the total number of actual disease-related gene sets found by the method in the same data. So the recall of the avelink-stat method was higher than others. There was only one overlapping GO term between the avelink-stat method and GSEA, and no overlapping GO term between the avelink-stat method and CORG-based method, NCFS-i and NCFS-c. However, there were seven overlapping GO terms between NCFS-i and NCFS-c. The results are shown in Table 8, the functional sets verified in the literature are marked with the letter 'a', and the overlapping gene sets are shown in bold.

Moreover, we performed literature verification for the significant functional gene sets obtained from the cent-stat method for AD. We verified 68 out of 116 (58.6%) gene sets in the literature. From the 23 sets in common with GSEA, 13 (56.5%) were verified; from the 93 sets that were different from GSEA, 55 (59.1%) were verified. From GSEA, 22 out of 42 gene sets were verified in the literature. The recall of the cent-stat method was significantly higher than the avelink-stat method and GSEA. In other words, the cent-stat method was more sensitive than the avelink-stat method and GSEA. We obtained a significantly higher precision value using our proposed methods (0.90 and 0.59) than GSEA (0.52). Particularly, we obtained a precision of 1 using a combination of the avelink-stat method and the cent-stat method. Additional functional gene sets that were not verified in the literature may be associated with AD. For example, prior studies found that 'actin binding',[63] 'actin cytoskeleton organization and biogenesis',[64] 'actin filament binding'[63] and 'actin filament organization'[65] were related to AD, and we identified two additional functional gene sets, 'actin filament based process' and 'actin filament bundle formation'. From the relationships between these sets, we believed that the two newly discovered functional sets were related to AD. This example demonstrated the effectiveness of these methods. The detailed results of this analysis are shown in Supplementary Table 2.

Using the NSCLC I data, Ruprecht Kuner[25] found 23 significant functional GO terms using GOstat[66] at a significance level of $P<0.0001$. From these 23 GO terms, 10, 11, 7, 5, 8 and 9 terms were detected by the avelink-stat method, the cent-stat method, GSEA, the CORG-based method, the NCFS-i method and the NCFS-c method, respectively. Our proposed methods yielded more verified disease-related gene sets than other methods, even if we obtained more significant gene sets using the CORG-based, NCFS-i and NCFS-c methods. We obtained a significantly higher precision value using a combination of the avelink-stat method and the cent-stat method (0.15) than other methods (<0.11). In addition, we obtained significantly higher recall values using our proposed methods (>0.43) than other methods (<0.39). Specifically, the term 'cell junctions' that was previously described[25] was detected by both of our methods but was not found using GSEA, the CORG-based method or the NCFS-i method. These results are shown in Table 6.

Computational complexity and repeatability

In the gene resampling-based permutation methods, gene resampling is the most time-consuming step. Suppose that we randomly resample m_i genes from a gene expression profile including m genes, calculate the distances between the two groups with different phenotypes and repeat this procedure n_{per} times. The computational complexity of these steps is approximately O ($N_{\text{per}}mm_i$). Even when n_{per} and m are large, the computational time is still acceptable.

The repeatability of the method is very important when detecting disease-related gene sets. We used the overlap coefficient and correlation coefficient between the two NSCLC data sets to describe the repeatability. These results are shown in Table 6 and Table 7b and c. We obtained a greater overlap coefficient while considering a combination of the avelink-stat and cent-stat methods (85.3%) compared with GSEA (77.8%) and the three other methods (a maximum of 64.4%). The Pearson's and Spearman correlation coefficients between the two data sets using the avelink-stat method (0.4290 and 0.4421, respectively) and the cent-stat method (0.4365 and 0.4785, respectively) were significantly larger than that with GSEA (0.0485 and 0.1326, respectively) and the CORG-based, NCFS-i and NCFS-c methods (Pearson's correlation coefficients of 0.4316, 0.4028 and 0.3844, and Spearman correlation coefficients of 0.4015, 0.3957 and 0.3731, respectively). These results confirmed that our proposed methods had better repeatability than other methods.

CONCLUSION

Compared with the IGA methods, we do not need to set a threshold for expression difference to classify genes between the case and control samples. In addition, we use the original gene data and not a summary statistic during analysis, whereas most GSA methods use summary statistics, such as the rank statistic used by GSEA. Therefore, this method fully utilizes each level of the gene expression data and is better able to detect functional sets, especially for gene sets containing more minor-effect genes. By analyzing simulated and real data, we determined that the label swapping-based permutation method is too sensitive and may result in high false-positives during the detection of significant disease-related functional gene sets; by evaluating precision, recall and the F_1 scores in the simulated data, we believe that the gene resampling-based permutation method is more suitable for gene set analyses. For the gene resampling-based permutation method, we determine that the statistical distance method is robust for the dimensions of the gene expression values.

Compared with GSEA and the CORG-based, NCFS-i and NCFS-c methods in the simulation experiment, we find that both the cent-stat and avelink-stat methods perform best, especially in detecting the disease-related gene sets containing more minor-effect genes.

When using the real data sets, we find that the cent-stat method is more sensitive than the avelink-stat method and other methods. The precision obtained from a combination of the avelink-stat and cent-stat methods is higher than the precision of other methods. Through validation using duplicate data sets, we determine that the repeatability of a combination of these two methods is better than other methods. Therefore, we recommend the use of the cent-stat method for the identification of more functional gene sets and a combination of these two methods for the more accurate identification of disease-related functional gene sets.

In this study, we perform distance-based gene set analysis with strong feasibility and effectiveness using GO as examples. In addition, we can analyze other functional sets, such as KEGG pathways and motif gene sets. We have prepared a simple and publically available Perl and R package for the centroid statistical distance method and the average linkage statistical distance method (http://bioinfo.hrbmu.edu.cn/dbgsa or http://cran.r-project.org/web/packages/DBGSA/).

ACKNOWLEDGEMENTS

This work was partially supported by the National Natural Science Foundation of China (grant nos. 30871394, 61073136, 91029717, 81172842, 60932008 and 61172098), the Specialized Research Fund for the Doctoral Program of Higher Education of China (grant no. 20112302110040), the Fundamental Research Funds for the Central Universities (grant no. HIT.ICRST.2010 022) and the Fund of Heilongjiang Health Department (grant nos.2011-204 and 2011-251).

1 Nam, D. & Kim, S. Y. Gene-set approach for expression pattern analysis. *Brief. Bioinform.* **9,** 189–197 (2008).

2 Emmert-Streib, F. & Glazko, G. V. Pathway analysis of expression data: deciphering functional building blocks of complex diseases. *PLoS Comput. Biol.* **7,** e1002053 (2011).

3 Hung, J. H., Yang, T. H., Hu, Z., Weng, Z. & Delisi, C. Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief Bioinform* **13,** 281–291 (2012).

4 Khatri, P. & Draghici, S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* **21,** 3587–3595 (2005).

5 Rivals, I., Personnaz, L., Taing, L. & Potier, M. C. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* **23,** 401–407 (2007).

6 Dopazo, J. Functional interpretation of microarray experiments. *OMICS* **10,** 398–410 (2006).

7 Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34,** 267–273 (2003).

8 Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).

9 Goeman, J. J., van de Geer, S. A., de Kort, F. & van Houwelingen, H. C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20,** 93–99 (2004).

10 Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S. *et al.* Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinform.* **8,** 242 (2007).

11 Hummel, M., Meister, R. & Mansmann, U. GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics* **24,** 78–85 (2008).

12 Nam, D., Kim, S. B., Kim, S. K., Yang, S., Kim, S. Y. & Chu, I. S. ADGO: analysis of differentially expressed gene sets using composite GO annotation. *Bioinformatics* **22,** 2249–2253 (2006).

13 Chi, S. M., Kim, J., Kim, S. Y. & Nam, D. ADGO 2.0: interpreting microarray data and list of genes using composite annotations. *Nucleic Acids Res* **39,** W302–W306 (2011).

14 Isci, S., Ozturk, C., Jones, J. & Otu, H. H. Pathway analysis of high-throughput biological data within a Bayesian network framework. *Bioinformatics* **27,** 1667–1674 (2011).

15 Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S. & Park, P. J. Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. USA* **102,** 13544–13549 (2005).

16 Goeman, J. J. & Buhlmann, P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23,** 980–987 (2007).

17 Ye, C. & Eskin, E. Discovering tightly regulated and differentially expressed gene sets in whole genome expression data. *Bioinformatics* **23,** e84–e90 (2007).

18 Thomson, G. An overview of the genetic analysis of complex diseases, with reference to type 1 diabetes. Best practice & research Clinical endocrinology & metabolism. *Best Pract. Res. Clin. Endocrinol. Metab* **15,** 265–277 (2001) [Research Support, US Govt PHS Review].

19 Scott, W. K., Pericak-Vance, M. A. & Haines, J. L. Genetic analysis of complex diseases. *Science* **275,** 1327–1330 (1997).

20 Lee, E., Chuang, H. Y., Kim, J. W., Ideker, T. & Lee, D. Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol* **4,** e1000217 (2008).

21 Sootanan, P., Prom-on, S., Meechai, A. & Chan, J. Pathway-based microarray analysis for robust disease classification. *Neural Comput. Appl.* **21,** 649–660 (2012).

22 Chan, J. H., Sootanan, P. & Larpeampaisarl, P. Feature selection of pathway markers for microarray-based disease classification using negatively correlated feature sets. *2011 International Joint Conference on Neural Networks (IJCNN 2011) IEEE*. p 3293–3299 (2011).

23 Thomas, R., de la Torre, L., Chang, X. & Mehrotra, S. Validation and characterization of DNA microarray gene expression data distribution and associated moments. *BMC Bioinform.* **11,** 576 (2010).

24 Webster, J. A., Gibbs, J. R., Clarke, J., Ray, M., Zhang, W., Holmans, P. *et al.* Genetic control of human brain transcript expression in Alzheimer disease. *Am. J. Hum. Genet.* **84,** 445–458 (2009) [Research Support, N.I.H., Extramural].

25 Kuner, R., Muley, T., Meister, M., Ruschhaupt, M., Buness, A., Xu, E. C. *et al.* Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung Cancer* **63,** 32–38 (2009).

26 Hall, J. S., Leong, H. S., Armenoult, L. S., Newton, G. E., Valentine, H. R., Irlam, J. J. *et al.* Exon-array profiling unlocks clinically and biologically relevant gene signatures from formalin-fixed paraffin-embedded tumour samples. *Br. J. Cancer* **104,** 971–981 (2011).

27 Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000) [Research Support, Non-US Gov't Research Support, US Gov't, PHS].

28 Hjorth, J. S. U. *Computer intensive statistical methods validation model selection and bootstrap* (Chapman and Hall, London, 1994).

29 Manly, B. F. J. *Randomization, bootstrap, and Monte Carlo methods in biology.* 3rd ed. (Chapman & Hall/CRC, Boca Raton, Fla.: London, 2007).

30 Zieffler, A., Harring, J. & Long, J. D. *Comparing groups randomization and bootstrap methods using R.* (Wiley-Blackwell, Oxford, 2011).

31 Strimmer, K. Fdrtool a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics* **24,** 1461–1462 (2008).

32 Chapman, S. *SimMetrics: Open Source Similarity Measure Library* (2007). Available from: http://sourceforge.net/projects/simmetrics/.

33 Chapman, S. *String similarity metrics for information integration* (2007). Available from: http://www.dcs.shef.ac.uk/~sam/stringmetrics.html.

34 Mecocci, P., MacGarvey, U. & Beal, M. F. Oxidative damage to mitochondrial DNA is increased in Alzheimer's disease. *Ann. Neurol.* **36,** 747–751 (1994).

35 Cordell, B. beta-Amyloid formation as a potential therapeutic target for Alzheimer's disease. *Annu. Rev. Pharmacol. Toxicol.* **34,** 69–89 (1994).

36 Lipton, S. A., Gu, Z. & Nakamura, T. Inflammatory mediators leading to protein misfolding and uncompetitive/fast off-rate drug therapy for neurodegenerative disorders. *Int. Rev. Neurobiol.* **82,** 1–27 (2007).

37 Tabner, B. J., Turnbull, S., El-Agnaf, O. & Allsop, D. Production of reactive oxygen species from aggregating proteins implicated in Alzheimer's disease, Parkinson's disease and other neurodegenerative diseases. *Curr. Top Med. Chem.* **1,** 507–517 (2001).

38 Datta, K., Sinha, S. & Chattopadhyay, P. Reactive oxygen species in health and disease. *Natl Med. J. India* **13,** 304–310 (2000).

39 Perry, G., Kawai, M., Tabaton, M., Onorato, M., Mulvihill, P., Richey, P. *et al.* Neuropil threads of Alzheimer's disease show a marked alteration of the normal cytoskeleton. *J. Neurosci.* **11,** 1748–1755 (1991).

40 Bamburg, J. R. & Wiggan, O. P. ADF/cofilin and actin dynamics in disease. *Trends Cell Biol.* **12,** 598–605 (2002).

41 Matus, S., Lisbona, F., Torres, M., Leon, C., Thielen, P. & Hetz, C. The stress rheostat: an interplay between the unfolded protein response (UPR) and autophagy in neurodegeneration. *Curr. Mol. Med.* **8,** 157–172 (2008).

42 Barnham, K. J., McKinstry, W. J., Multhaup, G., Galatis, D., Morton, C. J., Curtain, C. C. *et al.* Structure of the Alzheimer's disease amyloid precursor protein copper binding domain. A regulator of neuronal copper homeostasis. *J. Biol. Chem.* **278,** 17401–17407 (2003).

43 Lin, C. L., Bristol, L. A., Jin, L., Dykes-Hoberg, M., Crawford, T., Clawson, L. *et al.* Aberrant RNA processing in a neurodegenerative disease: the cause for absent EAAT2, a glutamate transporter, in amyotrophic lateral sclerosis. *Neuron* **20,** 589–602 (1998).

44 Cook, D. G., Forman, M. S., Sung, J. C., Leight, S., Kolson, D. L., Iwatsubo, T. *et al.* Alzheimer's A beta (1-42) is generated in the endoplasmic reticulum/intermediate compartment of NT2N cells. *Nat. Med.* **3,** 1021–1023 (1997).

45 Ebneth, A., Godemann, R., Stamer, K., Illenberger, S., Trinczek, B. & Mandelkow, E. Overexpression of tau protein inhibits kinesin-dependent trafficking of vesicles, mitochondria, and endoplasmic reticulum: implications for Alzheimer's disease. *J. Cell. Biol.* **143,** 777–794 (1998).

46 Nixon, R. A., Wegiel, J., Kumar, A., Yu, W. H., Peterhoff, C., Cataldo, A. *et al.* Extensive involvement of autophagy in Alzheimer disease: an immuno-electron microscopy study. *J. Neuropathol. Exp. Neurol.* **64,** 113–122 (2005).

47 Wallace, D. C. Mitochondrial diseases in man and mouse. *Science* **283,** 1482–1488 (1999).

48 Geula, C., Greenberg, B. D. & Mesulam, M. M. Cholinesterase activity in the plaques, tangles and angiopathy of Alzheimer's disease does not emanate from amyloid. *Brain Res.* **644,** 327–330 (1994).

49 Cutler, R. G., Kelly, J., Storie, K., Pedersen, W. A., Tammara, A., Hatanpaa, K. *et al.* Involvement of oxidative stress-induced abnormalities in ceramide and cholesterol metabolism in brain aging and Alzheimer's disease. *Proc. Natl Acad. Sci. USA* **101,** 2070–2075 (2004).

50 Stadelmann, C., Deckwerth, T. L., Srinivasan, A., Bancher, C., Bruck, W., Jellinger, K. *et al.* Activation of caspase-3 in single neurons and autophagic granules of granulovacuolar degeneration in Alzheimer's disease. Evidence for apoptotic cell death. *Am. J. Pathol.* **155,** 1459–1466 (1999).

51 Cassarino, D. S., Swerdlow, R. H., Parks, J. K., Parker, Jr. W. D. & Bennett, Jr. J. P. Cyclosporin A increases resting mitochondrial membrane potential in SY5Y cells and reverses the depressed mitochondrial membrane potential of Alzheimer's disease cybrids. *Biochem. Biophys. Res. Commun.* **248,** 168–173 (1998).

52 Toide, K., Okamiya, K., Iwamoto, Y. & Kato, T. Effect of a novel prolyl endopeptidase inhibitor, JTP-4819, on prolyl endopeptidase activity and substance P- and arginine-vasopressin-like immunoreactivity in the brains of aged rats. *J. Neurochem.* **65,** 234–240 (1995).

53 Connell, C. M., Shaw, B. A., Holmes, S. B., Hudson, M. L., Derry, H. A. & Strecher, V. J. The development of an Alzheimer's disease channel for the Michigan Interactive Health Kiosk Project. *J. Health Commun.* **8,** 11–22 (2003).

54 Kaminska, J., Hoffman-Sommer, M. & Plachta, M. The p24 family proteins–regulators of vesicular trafficking. *Postepy Biochem.* **56,** 75–82 (2010).

55 Ross, B. M., Moszczynska, A., Erlich, J. & Kish, S. J. Phospholipid-metabolizing enzymes in Alzheimer's disease: increased lysophospholipid acyltransferase activity and decreased phospholipase A2 activity. *J. Neurochem.* **70,** 786–793 (1998).

56 Lee, J. M., Calkins, M. J., Chan, K., Kan, Y. W. & Johnson, J. A. Identification of the NF-E2-related factor-2-dependent genes conferring protection against oxidative stress in primary cortical astrocytes using oligonucleotide microarray analysis. *J. Biol. Chem.* **278,** 12029–12038 (2003).

57 Hirai, K., Aliev, G., Nunomura, A., Fujioka, H., Russell, R. L., Atwood, C. S. *et al.* Mitochondrial abnormalities in Alzheimer's disease. *J. Neurosci.* **21,** 3017–3023 (2001).

58 David, D. C., Ittner, L. M., Gehrig, P., Nergenau, D., Shepherd, C., Halliday, G. *et al.* Beta-amyloid treatment of two complementary P301L tau-expressing Alzheimer's disease models reveals similar deregulated cellular processes. *Proteomics* **6,** 6566–6577 (2006).

59 Perry, T. L., Yong, V. W., Bergeron, C., Hansen, S. & Jones, K. Amino acids, glutathione, and glutathione transferase activity in the brains of patients with Alzheimer's disease. *Ann. Neurol.* **21,** 331–336 (1987) [Research Support, Non-US Gov't].

60 Brinton, R. D. Cellular and molecular mechanisms of estrogen regulation of memory function and neuroprotection against Alzheimer's disease: recent insights and remaining challenges. *Learn Mem.* **8,** 121–133 (2001).

61 Baloyannis, S. J. Mitochondrial alterations in Alzheimer's disease. *J. Alzheimers Dis.* **9,** 119–126 (2006).

62 Lukiw, W. J. & Bazan, N. G. Strong nuclear factor-kappaB-DNA binding parallels cyclooxygenase-2 gene transcription in aging and in sporadic Alzheimer's disease superior temporal lobe neocortex. *J. Neurosci. Res.* **53,** 583–592 (1998).

63 Harigaya, Y., Shoji, M., Shirao, T. & Hirai, S. Disappearance of actin-binding protein, drebrin, from hippocampal synapses in Alzheimer's disease. *J. Neurosci. Res.* **43,** 87–92 (1996).

64 Fulga, T. A., Elson-Schwab, I., Khurana, V., Steinhilb, M. L., Spires, T. L., Hyman, B. T. *et al.* Abnormal bundling and accumulation of F-actin mediates tau-induced neuronal degeneration *in vivo*. *Nat. Cell Biol.* **9,** 139–148 (2007).

65 Heredia, L., Helguera, P., de Olmos, S., Kedikian, G., Sola Vigo, F., LaFerla, F. *et al.* Phosphorylation of actin-depolymerizing factor/cofilin by LIM-kinase mediates amyloid beta-induced degeneration: a potential mechanism of neuronal dystrophy in Alzheimer's disease. *J. Neurosci.* **26,** 6533–6542 (2006).

66 Beissbarth, T. & Speed, T. P. GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics* **20,** 1464–1465 (2004).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)