Statistically invalid classification of high throughput gene expression data

Shahar Barbash & Hermona Soreq

The Edmond & Lily Safra Center for Brain Sciences and the Department of Biological Chemistry at the Hebrew University of Jerusalem.

Classification analysis based on high throughput data is a common feature in neuroscience and other fields of science, with a rapidly increasing impact on both basic biology and disease-related studies. The outcome of such classifications often serves to delineate novel biochemical mechanisms in health and disease states, identify new targets for therapeutic interference, and develop innovative diagnostic approaches. Given the importance of this type of studies, we screened 111 recently-published high-impact manuscripts involving classification analysis of gene expression, and found that 58 of them (53%) based their conclusions on a statistically invalid method which can lead to bias in a statistical sense (lower true classification accuracy than the reported classification accuracy). In this report we characterize the potential methodological error and its scope, investigate how it is influenced by different experimental parameters, and describe statistically valid methods for avoiding such classification mistakes.

Results

Defining the classification inaccuracy. The most common invalid protocol starts with the selection of a subset of transcripts, usually a couple hundred out of the (roughly) 25,000 human genes studied in different high throughput platforms, which are altered between the two test groups. It then uses this subset of genes for differentiating between the two groups. This approach usually yields a clear distinction between the two groups; however, this is not surprising given the fact that the genes on which the classification was based were initially selected BECAUSE of their differential expression between the two groups. The classification of such groups is often presented as one of the main findings of these studies; furthermore, some studies argue explicitly that, based on the clear distinction between groups, these genes may serve as a ‘signature’ or ‘fingerprint genes’ – biomarkers that elucidate the underlying mechanism of the studied disease and lead to diagnostic and drug discovery applications. The continuously reduced costs, improved technological accuracy, and shortened experimental time of high throughput gene expression studies predict a future increase in similar uses of high throughput platforms. However, many studies, including classification analysis, apply statistically invalid methods that can lead to erroneous conclusions.
Classifications were based on the 100 most drastically altered genes (Fig. 2a and d), which in both cases emerged as ‘successful’, presenting statistically significant (but nevertheless invalid) differences between the two tested groups. However, comparisons which tested 100 randomly-selected genes (Fig. 2b and e) and the full expression matrices (2c and f) both emerged as ‘failed’ for the experimental as well as the random datasets.

Of note, there are cases where comparison of full datasets yields significant, and hence valid, classification. Several methods are readily available which can correctly differentiate between groups based on high throughput datasets. 47% of the 111 studies we screened and which performed high-throughput classification analyses of gene expression datasets used one of these methods and avoided this statistical bias in classification experiments. Below we describe such valid classification methods, which are based on unbiased classification of the expression levels of all the tested genes, on a selection rational for identifying groups of candidate genes, or on splitting the cohort into training and test datasets.

The outcome of selecting the most drastically altered genes based on a statistical test performed per gene is that the variance between the members of a group becomes much smaller than the variance between members of different groups. This is shown in Figure 2g, where the decision regions for classification are based on the minimal distance between the coordinate and the mean of the groups in the plane composed of components one and two of the Principal Component Analysis (PCA) (see ‘methods’ for more information).

The first principal component which captures the maximal variance, and which is presented on the X axis in figure 2g, easily discriminates between the two groups. Although the results presented in figure 2g look ‘real’ or ‘significant’, the distribution of the total number of misclassifications (figure 2j-l) demonstrates that the outcome is always zero when using the 100 ‘most discriminative genes’. The chance of perfectly correct albeit arbitrary classification, for this scenario, is actually 100%. In other words, the P-value that should be assigned is 1.

Using statistical terminology the issue is one of selection (ascertainment) and errors in estimating effect sizes. Selecting a set of genes out of all genes based on a treatment or group differences when individual effects on gene expression are estimated with error leads to a correlation between treatment difference and random error. Creating a predictor out of the ‘best’ genes and fitting these back in the same data then amplifies these errors and may lead to serious bias.

Discussion

Several possible reasons for the high frequency of statistically biased classification studies are presented, as follows. First, while studies in
in the life sciences show a growing tendency to use computational tools, these are often internet-based software tools in which the user’s control over the applied parameters is very limited and the full picture of the underlying algorithm that served to develop this software tool is inaccessible to the typical user. Second, the seemingly unbiased automatic clustering of samples, which provides a reliable distinction between groups and presents a dramatic difference between different experimental conditions or patient groups. This points at the very attractive prospect of possible diagnostics or drug target applications. Of note, the conflict between the increasingly simple usage of large databases and the effort required for validation of single target genes can lead to limiting the number of validated genes that are selected for further focused studies. This, in turn leads to statistically biased selection of “signature” genes and is likely to lead to tedious, costly and frustrating efforts in translational studies.

Figure 2 | Comparison of classifications based on a subset of 100 'discriminative genes' (or numbers), 100 randomly-selected ones, and full matrices. Dendrogram plots are shown for groups of 6 patients and 6 healthy matched control volunteers, based on the 100 most altered genes (a), on 100 randomly-selected genes (b) and on the full expression matrices (c). (d–f) Dendrogram plots based on normally distributed random data of similar sizes. (g–i) Scatter PCA plots and decision regions based on the ‘minimal distance to group mean’ classifier, on the plane of PC1 and PC2. (j–l) Distributions of the number of misclassifications based on the ‘minimal distance to group mean’ classifier for the three protocols described in (a–c). Number of iterations = 500. See Methods for further details.

Figure 3 | Scope of invalid analyses in the literature. Stacked bar graphs represent the number of studies in which statistically invalid (red) or valid (blue) classification protocols were used, in the noted years.
We have shown that clustering based on a list of ‘informative genes’ presenting large expression differences between the studied groups leads to statistical bias and may thus draw biologically irrelevant classifications. However, several strategies exist for avoiding such errors and reaching valid clustering analyses. First, using an a priori list of genes for clustering tests, as done in, is absolutely essential, because the list would not depend on the specific experiment that is being assessed. For example, defining all of the genes involved in inflammation as a sub-set is a valid strategy for classification of groups with predictably distinct inflammatory features. Second, one can test for clustering based on the entire set of data, without any filtering, as done in. This method has actually been the preferred approach among those studies that avoided the statistically biased classification which we discuss here. Finally, one can divide the sample population into two sets: a training set in which the separating groups is assessed, as done in, and a test set, in which the clustering performance is assessed, as done in.

Recently, a paper by Nieuwenhuis et al. published in Nature Neuroscience showed that 50% of the articles examined made a statistical mistake by which the interaction between compared groups was not taken into consideration. In some of the examined articles the implication of this mistake was that the main claim of the article could not be statistically supported. We believe that the statistically biased classification we have drawn attention to is of similar significance, and that it may have a direct impact on scientific arguments that are made in various research fields.

We have shown that clustering based on a list of ‘informative genes’ presenting large expression differences between the studied groups leads to statistical bias and may thus draw biologically irrelevant classifications. However, several strategies exist for avoiding such errors and reaching valid clustering analyses. First, using an a priori list of genes for clustering tests, as done in, is absolutely essential, because the list would not depend on the specific experiment that is being assessed. For example, defining all of the genes involved in inflammation as a sub-set is a valid strategy for classification of groups with predictably distinct inflammatory features. Second, one can test for clustering based on the entire set of data, without any filtering, as done in. This method has actually been the preferred approach among those studies that avoided the statistically biased classification which we discuss here. Finally, one can divide the sample population into two sets: a training set in which the separating groups is assessed, as done in, and a test set, in which the clustering performance is assessed, as done in.

Recently, a paper by Nieuwenhuis et al. published in Nature Neuroscience showed that 50% of the articles examined made a statistical mistake by which the interaction between compared groups was not taken into consideration. In some of the examined articles the implication of this mistake was that the main claim of the article could not be statistically supported. We believe that the statistically biased classification we have drawn attention to is of similar significance, and that it may have a direct impact on scientific arguments that are made in various research fields.

**Methods**

**Dendrogram plot.** The nodes at the bottom of a dendrogram plot represent samples. While climbing up the dendrogram tree the nodes are joined together according to their proximity in a multidimensional space. In the beginning samples are scattered in a multidimensional space according to their expression values (of whole transcripts or subsets of it). In the first step, each pair of samples is joined together to form a new node whose coordinates are the means of the pair. In the second step a second level of nodes is formed similarly from the nodes in the previous step. This continues until all nodes have been assigned to a pair. The distance between each two nodes is the peak on the Y axis of the inverse U shaped line that connects them.

‘Minimal distance to group mean’ classifier. Decision regions are determined as follows: the samples are scattered on a two dimensional plane (PCA1 and PCA2 in Figure 2 g–l, but can be any other space as well). Each pixel on the plane is assigned as ‘red’ or ‘green’ according to its distance from the means of the groups. The color of the group with the closest mean to the pixel is the one that is assigned to it. Statistically biased classifications are counted as red dots in a green region, and vice versa.

**Acknowledgments**

The authors are grateful to Prof. Israel Nelken and Dr. Estelle R. Bennett, Jerusalem, for fruitful and comprehensive discussions and to the Gatsby, Legacy Heritage, ISF and Roetree foundations for support. S.B. is an incumbent of a PhD fellowship from the ELSC Brain Center.

**Author contributions**

H.S. and S.B. designed and wrote the main manuscript text. S.B. performed the simulations.

**Additional information**

**Competing financial interests:** The authors declare no competing financial interests.

**License:** This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/

**How to cite this article:** Barbash, S. & Soreq, H. Statistically invalid classification of high throughput gene expression data. Sci. Rep. 3, 1102; DOI:10.1038/srep01102 (2013).