BRIEFINGS IN BIOINFORMATICS. VOL 10. NO 5. 556–568 doi:10.1093/bib/bbp034

Stability and aggregation of ranked gene lists

Anne-Laure Boulesteix and Martin Slawski

Submitted: 13th May 2009; Received (in revised form): 5th June 2009

Abstract

Ranked gene lists are highly instable in the sense that similar measures of differential gene expression may yield very different rankings, and that a small change of the data set usually affects the obtained gene list considerably. Stability issues have long been under-considered in the literature, but they have grown to a hot topic in the last few years, perhaps as a consequence of the increasing skepticism on the reproducibility and clinical applicability of molecular research findings. In this article, we review existing approaches for the assessment of stability of ranked gene lists and the related problem of aggregation, give some practical recommendations, and warn against potential misuse of these methods. This overview is illustrated through an application to a recent leukemia data set using the freely available Bioconductor package GeneSelector.

Keywords: *Univariate analysis; differential expression; top-list; ranking; variability; bootstrap*

INTRODUCTION

Univariate analysis is an important part of most biomedical studies investigating high-dimensional molecular data. They yield so-called rankings, for instance, gene rankings in the case of microarray studies. Such rankings are displayed in almost all microarray-related biomedical publications, sometimes as the main research result, sometimes as a preliminary step to, e.g. more complex procedures like multivariate prediction of disease outcome, construction of genetic networks or combined analyses involving different types of -omics data. A multitude of ranking criteria have been proposed in the statistical and bioinformatic literature with the aim to cope with the 'small n large p problem' (where n is the number of observations and p the number of features). For example, the assessment of differential gene expression in the two-group setting has particularly attracted the attention of statisticians in recent years. Researchers typically compute a score such as the t-statistic reflecting the difference between both groups while taking the within-group variability into account. More sophisticated criteria which are particularly appropriate in the case of (very) small samples have been proposed in the literature, see recent comparison studies [1–3] for an overview of possible scores for the two-group setting.

From a practical point of view, the rank of a particular variable is often as important as the value of the statistic. Most often, the rank (not the value of the statistic) determines together with other aspects such as gene function whether the gene is selected for further analysis in future steps of the research project. For example, when there are 300 significant differentially expressed genes, the investigators will probably not test all of them extensively in further lab experiments. Conversely, they will probably examine the top-ranked genes very carefully even if no gene passes the multiple testing adjustment. Hence, providing a reliable list of top-ranking

Corresponding author. Anne-Laure Boulesteix, Department of Medical Informatics, Biometry and Epidemiology (IBE), Marchioninistr. 15, 81377 Munich, Germany. Tel: +49 89 7095-7598; Fax: +49 89 7095-7491; E-mail: boulesteix@ibe.med.uni-muenchen.de

Anne-Laure Boulesteix is an assistant professor for computational molecular medicine at the Faculty of Medicine of the University of Munich, Germany since 2009. She obtained her PhD in statistics in 2005 and then worked as a post-doc in medical statistics and bioinformatics. Her researches focus on the statistical analysis of high-dimensional molecular data, especially gene expression and SNP data.

Martin Slawski obtained his diploma in statistics in 2008 from the University of Munich, Germany. During his studies of statistics, he worked on various projects on high-dimensional data analysis. He is now a PhD student in the Machine Learning Group at the Department of Computer Science of the Saarland University, Saarbrücken, Germany.

genes is probably at least as important as improving statistical power of multiple testing procedures at any price. In this spirit, Mukherjee *et al*. [4] state that, 'a realistic goal [of the statistician or bioinformatician] is to narrow the field for further analysis, to give geneticists a short-list of genes which are worth investing hard-won funds into analysing', which Aerts *et al*. [5] formulate as a 'need for prioritization' in a slightly different context. A related problem going beyond the scope of this article is the adjustment for multiple testing [6–8]. Here, we focus on the ordering of the genes, not on the threshold of statistical significance.

Note that the concept of top-ranking variable may refer to other situations than the usual two-group scenario in gene expression studies. For instance, the phenotype of interest may be a quantitative trait instead of a binary outcome, or even a censored survival time. Categorical phenotypes with more than two categories are also common, and more difficult to handle than the two-group case. The concepts and ideas presented in this article restrict neither to the comparison of two groups nor to gene expression data, although the illustrating examples refer to such studies. Very generally, we denote as 'statistic' the criterion used to rank variables, where large values of the statistic correspond to top-ranking variables.

An important aspect of univariate analyses which is in our opinion under-considered in the literature is the variability of the obtained ordered lists. By variability, we mean (i) the difference between rankings obtained from the same data set based on different ranking criteria, and (ii) the variability of the lists obtained with a unique method but with slightly modified versions of the data set. Ranking methods should be stable in the sense that the list remains approximately the same even after small changes in the composition of the data sets or in the data preparation. Both biomedical and statistical studies tend to ignore these two sources of variability and simply consider the obtained ranked gene list as an unequivocal and definitive result.

In the present article, we systematically survey two important aspects of the variability of ranked lists: (i) the measurement of ranking stability, which is reviewed in 'Stability of gene rankings' section, and (ii) the aggregation of ranked lists in the hope to obtain a more reliable ranking, as introduced and discussed in 'Deriving more accurate ranking criteria through aggregation' section. Along with a systematic literature review on stability and aggregation in the context of ranked gene lists, we summarize the state-of-the-art within a unifying framework, which encompasses most of the existing procedures, we give some practical recommendations, and warn against potential misuse of methods for stability assessment and aggregation. The discussed methods are illustrated through an application to a recent leukemia microarray data set [9, 10] using the add-on Bioconductor package `GeneSelector` [11] in the statistical software environment `R`.

A UNIQUE GENE RANKING?

Notations

Although this is not always admitted in biomedical publications, the ranked lists obtained from univariate analyses should not be considered as fixed universal results. After giving a formal definition of the term 'ranking' as considered in the present article, we briefly review important sources of variations for gene rankings.

The term 'data set' denotes a pair $\mathcal{D} = (\boldsymbol{x}, \boldsymbol{y})$, where the $n \times p$ matrix $\boldsymbol{x} = (x_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,p}}$ contains n observations of the random vector $(X_1, \ldots, X_p)'$ (for instance, the expression levels of p genes), and $\boldsymbol{y} = (y_1, \ldots, y_n)$ stores either the response variable Y of interest for these n observations or, e.g. an experimental condition fixed by design. In this article, we define a *ranking* of the variables $X_1, \ldots, X_p$ as a permutation $\boldsymbol{r} = (r_j)_{j=1,\ldots,p}$ of $(1, \ldots, p)$, where r_j is the rank of the variable X_j with respect to its association with Y. A small rank indicates strong association between the considered gene and Y, either positive or negative. A ranking yields an ordered *list* $\boldsymbol{l} = (l_m)_{m=1,\ldots,p}$ defined by

$$l_m = j \;\Leftrightarrow\; r_j = m \quad \text{for all } j, m = 1, \ldots, p. \qquad (1)$$

For instance, in the case of differential gene expression, $r_{1786} = 1$ and $l_1 = 1786$ would mean that X_{1786} is identified as the most differentially expressed gene. If $\boldsymbol{l}$ is an ordered list, the k top genes $l_1, \ldots, l_k$ form the so-called top-k list (usually $k \ll p$). For example, biomedical articles often report top-20 or top-50 lists. Note that we do not consider rankings with ties for the sake of simplicity.

Different ranking criteria

While methodological articles presenting new statistical ranking methods often include a comparison study for demonstrating their superiority over

existing approaches, many biomedical publications show results obtained with a single statistic. However, different ranking statistics often result in very different ranked lists, even when considering top-ranking variables only. Methodological articles usually highlight these differences because they can be considered as arguments in favor or against a particular method. For instance, one often reads sentences like 'method A correctly identifies three well-known markers as differentially expressed, whereas these markers are not recognized as top-ranking by method B'. In contrast, differences between rankings are often ignored in biomedical publications presenting a new data set, because they would make the interpretation of results more confusing and raise doubts on the reliability of the data (for instance, a referee might argue that more experiments should be performed to make the list less ambiguous).

The multiplicity of methods and the question of how to deal with the different results that they produce are very general issues in all research fields. However, they are particularly relevant in the context of different expression studies in the '$n \ll p$' setting for two reasons. First, traditional approaches such as the standard *t*-test for two independent samples are usually considered as inappropriate in the case of very small sample sizes which are common in microarray settings, hence making the development of alternative approaches necessary. This strong need has been recognized by the research community and many teams have suggested their own method to provide more reliable rankings of differentially expressed genes, hence indirectly giving rise to another multiplicity problem. Second, univariate analyses in the small n large p scenario are well known to produce highly instable outputs, in the sense that a small change in the data or a minimal modification of the ranking criterion often results in a fully different ordering of the features.

Several tens of criteria for ranking genes according to their differential expression across two groups have been described in the literature in the last decade, making an exhaustive enumeration burdensome. The rest of this section gives a concise overview of the most widely used methods and briefly discusses their characteristics. Note that most of them can be interpreted as test statistics in the hypothesis testing framework. In this article, they are considered as ranking criteria rather than as test statistics, since we focus on the ordering of the genes rather than on the assessment of the statistical significance. The available ranking criteria can be roughly divided into the following broad categories [3]:

- Simple approaches such as the fold-change, the classical *t*-statistic with equal or unequal variance and Wilcoxon's rank sum statistic.
- *Ad hoc* modifications of the classical *t*-statistic such as the SAM statistic [12], which can also be interpreted as a Bayesian *t*-statistic where the denominator is stabilized by a global prior standard error, or Efron's 90% rule derived from a mixture model [13].
- Regularization methods, for instance, hierarchical Bayes methods performing regularization via prior distributions, where hyperparameters of the prior distribution are estimated from a large set of genes [14–17]. A regularized *t*-statistic can also be obtained through a James–Stein shrinkage procedure [3].

In the rest of this article, we illustrate the reviewed methods using the `R` statistical software (http://www.r-project.org) through an application to the leukemia data set available from the Bioconductor `ALL` package [9, 10]. The data set gives the expression levels of 12 265 genes for 95 T-cells and 33 B-cells samples.

The data are obtained from the `ALL` package with the `R` command lines:

```
R> library(ALL)
R> data(ALL)
R> X<-exprs(ALL)
R> y<-phenoData(ALL) [,4]
```

The `GeneSelector` add-on package [11] is publicly available from the `Bioconductor` platform (www.bioconductor.org) and can be installed and loaded using the commands

```
R> install.packages(''GeneSelector'')
R> library(GeneSelector)
```

A total of 15 established ranking criteria are either directly implemented in `GeneSelector` or called via wrapper functions from other packages. Given `X` and `y` as above, these functions are schematically called as `RankingCriterion(X, y, parameter.1,...,parameter.m)`. The arguments `parameter.1,...,parameter.m` may optionally be used to set method parameters to a different

value than the predefined default. Five examples including both standard methods (fold-change, ordinary *t*-statistic) and recently developed challengers (Limma [16], Fox-Dimmic *t*-statistic [17] and shrinkage *t*-statistic [3]) are given below.

```
R> fc <- RankingFC(X, y)
R> tstat <- RankingTstat(X, y)
R> limma <- RankingLimma (X, y, proportion = 0.01)
R> foxdimmic <- RankingFoxDimmic (X, y, m = 100)
R> shrinkt <- RankingShrinkageT (X, y)
```

The top-lists yielded by the ranking procedures can be easily displayed using the function `toplist`:

```
R> methodlist <- list(fc, tstat, limma, foxdimmic, shrinkt)
R> topten <- matrix(nrow = 10, ncol = length(methodlist))
R> for (i in seq(along = methodlist)) topten[, i] <- toplist(methodlist [[i]], top=10, show=FALSE)$index
R> colnames(topten) <- unlist (lapply(methodlist, slot, ''method''))
R> print(topten)
```

	Foldchange	ordinaryT	Limma	FoxDimmicT	ShrinkageT
[1,]	8399	8399	8399	8399	8399
[2,]	8173	8225	8225	8225	8321
[3,]	9478	3268	3268	8173	8064
[4,]	8172	5064	5064	9478	9478
[5,]	11 834	1174	1174	11 834	8225
[6,]	6702	7106	7106	2673	3268
[7,]	2673	8172	8172	8172	7414
[8,]	9932	8917	8917	7106	11 719
[9,]	8321	3067	3067	6702	1174
[10,]	9407	9034	8173	122	8917

While gene `8399` is unambiguously at the very top position, there are already three different candidates for the second position. Furthermore, except for gene `8399`, there is no gene present in all top-10 lists, as output by the method `GeneSelector`:

```
R> genesel <- GeneSelector(methodlist, maxrank = 10)
R> show(genesel)
GeneSelector run with gene rankings
  from the following statistics:
Foldchange
ordinaryT
Limma
FoxDimmicT
ShrinkageT
Number of genes below threshold rank 10
  in all statistics:1
R> which(slot(genesel, ''selected'') == 1) [1] 8399
```

The multiplicity of available methods may lead to a substantial publication bias in the sense that some researchers might choose their ranking statistic on the basis of its results. For example, if the 'favorite gene' is not identified as top-ranking by a particular criterion, researchers are likely to try out another statistic or even to develop a new assessment procedure that better corresponds to the expected or observed data structure. By favorite feature, we mean for instance a feature that is expected to be relevant based on biological knowledge, the feature that is expected to rank best based on a previous study, or conversely a feature which has not yet been identified as relevant in this context and may thus yield an innovative marker. The strategy consisting of choosing the ranking statistic *a posteriori* based on the obtained results should be banned, since it may yield a substantial optimistic bias and lead to 'wrong research findings' [18], as quantitatively assessed in the case of high-dimensional class prediction [19].

Different perturbed versions of the data set

When addressing the variability of gene lists, one can also consider the variations resulting from small changes in the data set. One may obtain a completely different ranked list, not only by using a different ranking statistic but also by considering subsets (or other modified versions) of the data set. Ideally, one would expect that the ranked list remains approximately the same even if, e.g. a few observations are removed from the data set. This kind of stability is often investigated through resampling techniques, which are briefly reviewed in the rest of this section together with two additional methods.

There are many different resampling variants whose theoretical properties are relatively well known, see for instance the textbooks [20, 21]. In the Jackknife procedure (also denoted as 'subsampling'), a number $d < n$ of observations are randomly selected and removed from the data set. When the number of distinct subsets of size d is moderate, they are often all considered successively, i.e. $B = n!/(d!(n-d)!)$. Otherwise, the subsets of observations

to be removed are generated randomly, with B usually depending on the induced computational expense. The special case 'leave-one-out' is obtained when one removes only $d=1$ observation at a time. An alternative to the Jackknife is the so-called 'bootstrap' method: one draws n observations out of n with replacement.

If the scale of the measurements is metric, perturbed data sets can also be generated by adding noise to the data matrix, for instance independent normally distributed terms with zero mean and 'small' (feature-specific) variance.

If Y takes values in a finite set of labels, and there is uncertainty about the correctness of label assignments, the consequences of erroneous labeling might be investigated by generating perturbed data sets by changing the label of a number d of randomly selected observations. Note that changing the labels usually affects the results more than simply removing the observations. In practice, d should thus be small.

In the context of univariate analyses, one can then repeat the variable ranking procedure for each modified version of the original data set successively and compare the resulting rankings by measuring their similarities. Such strategies have often been used to study the stability of gene rankings [22–29]. These methods are applied in medical research for empirically checking the stability of the identified genes [30, 31]. Bootstrap sampling and Jackknife are by far the most widely used procedures.

As an illustration, three of the perturbation schemes outlined above (Jackknife, bootstrap and label swap) are applied together with the Limma ranking procedure [16] using the `GeneSelector` package. As a preparatory step, the methods `GenerateFoldMatrix` and `GenerateBootMatrix` are called in order to specify how the perturbed data sets have to be generated. Note that the function `set.seed` is used here for the sake of reproducibility.

```
R> set.seed(1)
R> leave10out <- GenerateFoldMatrix (y
   = y, replicates = 50,
k = floor(length(y) * 0.1),
    minclassize = 20)
R> set.seed(2)
R> boot <- GenerateBootMatrix (y = y,
   replicates = 50,
maxties = 3, minclassize = 20)
```

The first function call randomly selects about 10% of all observations for 50 iterations, while the second one produces 50 bootstrap data sets. Note that the two methods `GenerateFoldMatrix` and `GenerateBootmatrix` allow resampling with constraints (as specified by the argument `minclassize` corresponding to the minimum number of observations from each class), which is often necessary in the case of very small samples. The chosen ranking procedure can then be applied to the perturbed data sets using the method `RepeatRanking`, for example:

```
R> limma_leave10out <- RepeatRanking
   (limma, leave10out, scheme = ''sub
   sampling'')
R> limma_change10 <- RepeatRanking
   (limma, leave10out, scheme =
   ''labelexchange'')
R> limma_boot <- RepeatRanking(limma,
   boot)
```

The Limma ranking is computed based on perturbed data sets obtained by removing 10% of the observations using the first command or by changing the label of these observations using the second command, while the third command computes the rankings based on bootstrap samples. Again, the method `toplist` allows to visualize the results: among others, it outputs the number of times each gene is placed at a particular top-position.

In the literature, such approaches have been used for two main purposes: (i) the assessment of the stability of variable rankings, and (ii) the derivation of aggregated rankings which are believed to be more reliable than rankings obtained from the original data set. Both approaches are reviewed in 'Stability of gene rankings' and 'Deriving more accurate ranking criteria through aggregation' sections, respectively.

STABILITY OF GENE RANKINGS

Framework

In this article, the term stability refers to the similarity between rankings obtained with different (similar) criteria (see 'Different ranking criteria' section) or based on different perturbed version of the data set (see 'Different perturbed versions of the data set' section). This section is devoted to the measurement of stability in terms of similarities between

lists in this context. Note that one may also investigate the stability of ranked lists across different platforms or across different studies within a meta-analysis. It is well known that results from different studies often show poor overlap, and that the combination of several studies in form of a meta-analysis may yield more reliable results. Many groups compare and combine results from, e.g. different platforms or different labs [32–34]. While some approaches are rank-based [33], most meta-analysis methods are based on the statistic itself and thus do not fit into the scope of the present article. A recent overview can be found in [35].

In our framework, we consider two rankings $\boldsymbol{r}$ and $\boldsymbol{r}'$ associated to two lists $\boldsymbol{l}$ and $\boldsymbol{l}'$ through Equation (1). A stability measure is a function s of $\boldsymbol{r}$, $\boldsymbol{r}'$ and possibly additional parameters that measures the similarity between the two lists. Reasonable stability measures focus on the top-ranking variables, because similarities between the ranks of relevant variables are more interesting than similarities between irrelevant variables. For instance, usual approaches may concentrate on top-k lists (for some $k \ll p$) or give more importance to the top of the list through appropriate weighting. Note that in practice the stability measures reviewed below often have to be accommodated to take more than two rankings into account. There are essentially two approaches to do that. Some criteria can be generalized in a way that they consider all rankings simultaneously, especially the criteria based on the concept of overlap [36]. The second approach consists of summarizing pairwise stability measures, for instance through averaging [29]. This can be done by considering all pairs of rankings successively [29], or by comparing a 'reference list' (for instance, the list derived from the original data set) to all the other [23].

Examples of stability measures

Many approaches are based on the comparison of sets consisting of a fixed number k of top-genes, e.g. $k = 100$. Common criteria for measuring the similarity between two gene lists are the size of the intersection of the two top-k lists

$$s^{(1)}(\boldsymbol{r}, \boldsymbol{r}', k) = \sum_{j=1}^{p} I(r_j \leq k \ \wedge \ r'_j \leq k),$$

where I denotes the indicator function [$I(A) = 1$ if A is true, $I(A) = 0$ otherwise], or the proportion $s^{(1)}(\boldsymbol{r}, \boldsymbol{r}', k)/k$ of genes in the top-k list from $\boldsymbol{l}$ that are also in the top-k list from $\boldsymbol{l}'$, also denoted as percentage of overlap or percentage of overlapping genes (POG) [37]. Such stability measures are used by numerous researchers [28, 29, 38, 39].

In the same vein, Irizarry *et al.* [40] suggest a visualization technique called 'CAT plot' (standing for 'Correspondence At the Top') that represents the proportion of overlap of the top-k lists versus k. A variant can be obtained by dividing the size of the intersection by the size of the union, yielding

$$s^{(2)}(\boldsymbol{r}, \boldsymbol{r}', k) = \frac{s^{(1)}(\boldsymbol{r}, \boldsymbol{r}', k)}{2k - s^{(1)}(\boldsymbol{r}, \boldsymbol{r}', k)}.$$

The stability measure $s^{(2)}(\boldsymbol{r}, \boldsymbol{r}', k)$ is often used in practice [36, 39]. It can be seen as a special case of Jaccard's index obtained when the two compared sets are of the same size k. For all these methods, there is inevitably some arbitrariness in the choice of the cutoff parameter k. In practice, one might consider several different arbitrarily chosen numbers of genes (e.g. $k = 50$, 100, 200) [2]. An other pitfall of these methods is that they follow the nothing-or-all principle. All genes in the subset are given the same weight independently of their rank, whereas the other are completely ignored. The major advantage of such approaches is their simple interpretation in terms of intersection.

The so-called overlap score [41] implemented in the package `OrderedList` [42] is also related to this class of stability measures, except that it considers similarities between the genes from both the top and the bottom of the lists, corresponding to, e.g. up- and downregulation. A pitfall of this approach is that the list is considered as symmetric with the top- and bottom-genes being implicitly equally relevant. If, as assumed in the present article, only the top-genes are important (i.e. if both up- and downregulated appear at the top of the list in the two-group setting), the score can be reformulated as

$$s^{(3)}(\boldsymbol{r}, \boldsymbol{r}', \alpha) = \sum_{k=1}^{p} w_\alpha^{(k)} \cdot s^{(1)}(\boldsymbol{r}, \boldsymbol{r}', k),$$

where $w_\alpha^{(k)}$ $(k = 1, \ldots, p)$ are weights decreasing with increasing k. In this representation, the score is a weighted combination of the sizes of the intersections of the two top-k lists for different k values. It avoids the difficult choice of k but involves an additional parameter α determining the weights $w_\alpha^{(k)}$. Yang *et al.* [41] suggest to choose exponential weights of the form $w_\alpha^{(k)} = e^{-\alpha k}$ in order to put more

weight on the top of the list. The parameter α determines the depth of the top-list that is considered as relevant for the comparison of the two lists. They suggest a procedure for choosing the optimal value of α: the score is computed for different values of α, both for subsamples as well as for subsamples with permuted class labels. The optimal value of α is then the value that best separates permuted subsamples from subsamples with true class labels in receiver operating characteristic (ROC) curve analysis. To our knowledge, this approach is the only one that determines the 'depth of the comparison' based on an objective criterion.

Other criteria which can also be formalized in a similar manner are the overlapping probability of top ranking gene lists derived based on the hypergeometric distribution [43], Dice–Sorensen's index and Ochiai's index [39], or simply the size of the union [29]. Methods based on the hypergeometric distribution have the advantage that the associated p-value provides a natural choice for the cutoff k as demonstrated in a different context [44, 45]. Note that it is also usual to consider top-lists of different sizes, for instance when the results of different studies on the same disease are compared in the context of meta-analysis. As an example, the so-called recovery score [23] considers lists of genes with p-values falling below a given threshold rather than top-k lists. This method is thus not purely rank-based, although it can be used in a similar way as the methods described above.

Set-based stability measures can be computed using the method `GetStabilityOverlap` from the `GeneSelector` package. For example, the stability of the Limma ranking procedure when 10% of the observations are removed from the data can be assessed through

```
R> limma_leave10out_setbased <- GetStabilityOverlap(limma_leave10out,
scheme = ''original'', decay =
      ''exponential'', alpha = 0.16)
R> summary(limma_leave10out_setbased,
   measure = ''intersection'', position = 10)

summary of intersection counts (with
  respect to reference data set):
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.800 0.900 0.900  0.938 1.000 1.000
expected score in the case of
  no-information: 0.0007920792
R> summary(limma_leave10out_setbased,
   measure = ''overlap'', position = 10)

summary of overlap scores (with
  respect to reference data set):
Min.  1st Qu. Median Mean 3rd Qu. Max.
0.863 0.916  0.940  0.935 0.959 1.000
expected score in the case of
  no-information: 0.0003353366
```

The first summary command displays the mean percentage of overlap with $k = 10$, while the second one shows the corresponding mean overlap score. As specified through the argument `scheme = ''original''`, similarity is measured between the ranking from the original data set and the rankings from perturbed data sets (and not between all pairs of perturbed data sets). The graphics depicted in Figure 1 can be obtained by applying the `plot` method to the objects output by `RepeatRanking` and `GetStabilityOverlap`, respectively. We conclude that the removal of only 10% of the observations modifies the top-10 list moderately, with a percentage of overlap of 85% or more. It can also be seen from Figure 1 that the percentage of overlap and the overlap score remain approximately constant when k increases. However, they strongly depend on the proportion of removed observations: for example, the mean intersection for $k = 10$ decreases to approximately 0.79 when 50% of the observations are removed, or when bootstrap samples are used. Note that the stability also depends on the ranking method.

Adaptations of well-known distance measures such as, e.g. Spearman's or Kendall's correlation form a distinct family of methods. They essentially consider the distance between the two considered rankings $\boldsymbol{r}$ and $\boldsymbol{r}'$, while focusing on the top of the lists. The definition of the term 'distance' depends on the particular method. For example, existing methods that have been used in bioinformatics are the 'top-k list Kendall distance' [33] and Spearman's footrule based on the set of the genes appearing in at least one top-k list [46]. The reference textbook [47] discusses further metric methods for analyzing partially ranked data consisting of several top-k lists.

An other important example used in the context of gene lists is a weighted version of Spearman's footrule denoted as Canberra distance [29], which is defined as $\sum_{j=1}^{p} |r_j - r'_j|/(r_j + r'_j)$

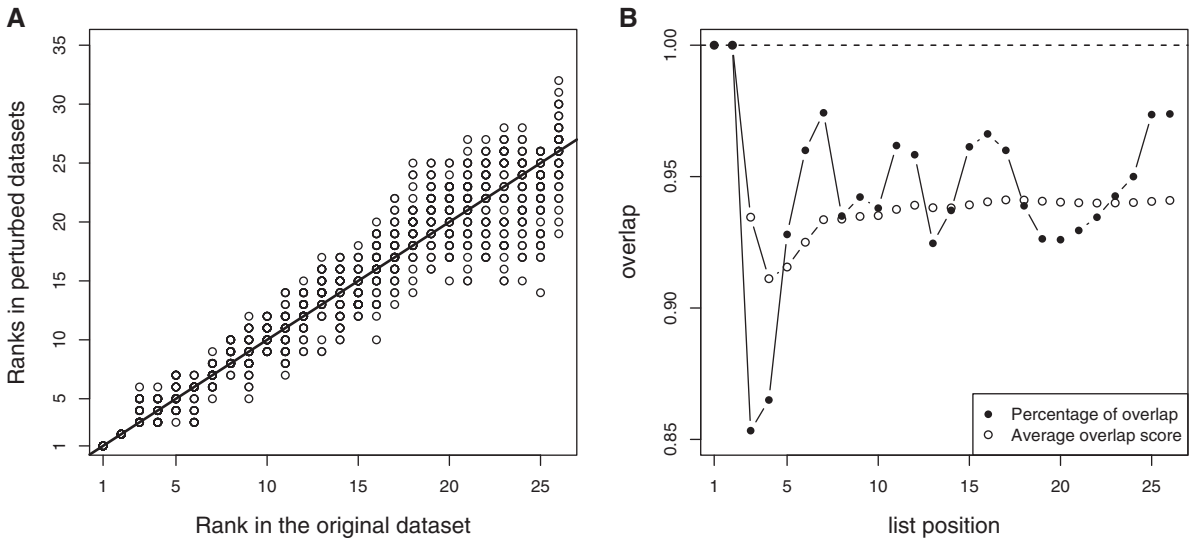

Figure 1: (**A**) Rank in the perturbed data sets (10% left out) against rank in the original data set. (**B**) Average percentage of overlap and average overlap score against the list position k.

and gives more weight to the top of the list through the denominator $r_j + r'_j$. The related Canberra distance $\sum_{j=1}^{p} |\min(r_j, k+1) - \min(r'_j, k+1)| / (\min\{r_j, k+1\} + \min\{r'_j, k+1\})$ with location parameter $k+1$ focuses on the top-k list. Some interesting properties can be derived in the group theoretical framework [47], for instance the mean in the noninformation case [29].

Remarks on stability

It seems that, in practice, two gene lists may show apparently poor overlap although each list comprises mostly truly differentially expressed genes [48]. Based on this idea, an alternative stability criterion accounting for correlation between genes was recently proposed [37]. Roughly, a variant of the percentage of overlap is computed by counting both the genes shared by the two top-lists and the genes from one of the top-lists that are significantly correlated with at least one gene from the other top-list. Based on this measure, the authors conclude that gene lists are often not as different as they might seem at the first glance. Note, however, that this method is not purely rank-based since it uses additional information on correlations.

Beside the critical assessment of the results' reliability, there are other motivations for assessing the stability of univariate rankings. One of these motivations is the optimized choice of a ranking criterion or of the parameters involved in a given ranking criterion. Several variants of this approach are proposed by different groups of authors [49, 50]. They suggest to 'learn' the ranking criterion based on the reproducibility of the ranked lists obtained from bootstrap samples, i.e. to select the parameter value or the ranking criterion yielding the highest stability. However, we point out that, although stability should be an important criterion for choosing a particular statistic, the choice should not be based solely on the stability, because a stable ranking procedure does not necessarily rank the genes properly. This becomes obvious if we consider a ranking criterion that always assigns each gene to the same rank independently of the sample. Such a ranking would obviously be stable, but fully inappropriate. In this vein, the so-called fold-change criterion (which considers the mean difference between the two groups after log-transformation of gene expression data) is often found to yield higher stability than criteria involving some variance estimation, but it is also well-known that it may not identify the right genes. In a nutshell, an ideal ranking criterion has both a low variance (i.e. a high stability) and a low bias.

The second motivation for assessing the stability of ranking procedures is that stability provides an objective criterion for computing the minimal sample size required in an experiment. In this spirit, one can determine whether the sample size is sufficient based on the stability of the obtained ranking [23]. Using this method, these authors found that a stability plateau is reached for sample sizes between $n = 8$ and $n = 15$ with the real data sets that they investigated. They argue that stability-based determination of the minimal sample size is more

realistic than model-based methods which rely on the quality of the data generating model—the power being unknown in real data studies. The 'probability of selection' method [22] is based on a similar idea. In this framework, the power is defined as the probability for a gene with true rank $\leq k_1$ to be ranked among the first k_0 genes, where k_0 and k_1 are adequately chosen integers. The power is assessed for different sample sizes in data-driven simulations.

DERIVING MORE ACCURATE RANKING CRITERIA THROUGH AGGREGATION

Framework

An issue related to the stability and similarity of gene rankings is the derivation of aggregated rankings in the hope that they will be more reliable than the original ranking obtained using the original data set. The term 'reliability' is of course difficult to define in the context of gene rankings, since the truth is not known. In practice, reliability is often assessed based on simulated data, which we consider as suboptimal since this approach usually oversimplifies the data structure, or by comparing the obtained top-genes to previous biological knowledge. Alternatively, some authors assess their lists of genes within a supervised learning framework, i.e. by employing them for prediction and computing the resulting error rate using an appropriate evaluation design [51]. In the slightly different context of variable selection for regression, the advantages of repeating the procedure on a large number of subsamples is well documented [52], with nice consistency properties.

Formally, an aggregated ranking or an aggregated list are obtained by 'summarizing' a collection of rankings $\{\boldsymbol{r}^{(b)},\ b = 1, \ldots, B\}$ or a collection of lists $\{\boldsymbol{l}^{(b)},\ b = 1, \ldots, B\}$, respectively. Note that this framework is very general. A collection of rankings/lists can be obtained both from a collection of perturbed data sets for a fixed ranking procedure or, conversely, from applying different ranking procedures to a fixed data set. Meta-analysis approaches that combine results from different studies using ranked-based procedures also fit in this framework.

Examples of aggregation procedures

A straightforward method to combine several rankings consists of computing a univariate summary statistic for each gene j, e.g. the mean $\frac{1}{B}\sum_{b=1}^{B} r_{jb}$ or a quantile of r_{jb}, $b = 1, \ldots, B$. The genes are then reranked according to their summary statistic. In particular, high quantiles (for instance, 0.75, 0.9 and 0.95) may be relevant because they favor genes which are never badly ranked, hence potentially eliminating genes whose apparent good rank is due to a single outlying observation.

Another approach focuses on the frequency of selection of the considered gene j within the top-k list over different rankings $\boldsymbol{r}^{(1)}, \ldots, \boldsymbol{r}^{(B)}$, also denoted as the 'selection probability function' [22]. The idea of selecting genes based on their frequency of selection among a large number of perturbed data sets is also proposed in several similar variants by numerous independent researcher groups [28, 29, 53]. A major pitfall of all these approaches is that they heavily depend on the arbitrarily fixed threshold k. By trying several thresholds successively, one would obtain substantially different rankings. Another problem is that they ignore the rank of the selected genes within the considered subset. As outlined in 'Stability of gene rankings' section in the context of stability assessment, they follow the all-or-nothing principle. However, their simplicity (they are easy to interpret) and their versatility (they can also be used to measure stability) are major advantages.

A more complex approach is based on Markov chain models and inspired from webpage ranking methodology [33]. Given a collection of rankings $\boldsymbol{r}^{(1)}, \ldots, \boldsymbol{r}^{(B)}$, one extracts all genes j with rank not greater than a threshold k in at least one of the rankings, i.e. one considers the set $\mathcal{E}_k = \{j : \exists b \in \{1, \ldots, B\} : r_{jb} \leq k\}$. Each gene $j \in \mathcal{E}_k$ represents one state of a Markov chain, with $\mathcal{E}_k$ being the state space. The event $r_{jb} < r_{j'b}$ is interpreted as a transition from state j' to state j. The corresponding transition matrix of the Markov chain is estimated based on the frequencies of the events $\{r_{jb} < r_{j'b}$ (for $j, j' \in \mathcal{E}_k$) in the collection of rankings $\boldsymbol{r}^{(1)}, \ldots, \boldsymbol{r}^{(B)}$. The obtained estimate is modified to achieve ergodicity, which guarantees the existence of a unique stationary distribution of the Markov chain. An aggregated ranking is then obtained by ordering the genes decreasingly according to their stationary probabilities.

Another recently proposed procedure that uses the sets $\mathcal{E}_k$ is based on the cross-entropy Monte Carlo optimization technique [46]. It searches iteratively for the optimal list in terms of the minimization of the sum of weighted distances between the

candidate aggregated list and each of the input lists. Two different distance measures are considered: Kendall's tau and Spearman's footrule. An advantage of this method is that it can put different weights to each input list depending on its importance/reliability. Note that the two approaches based on $\mathcal{E}_k$ yield aggregated rankings for the top-genes only.

Aggregation can be simply performed using the `GeneSelector` package:

```
R> methodRR<-MergeMethods(methodlist)
R> agg_mean <- AggregateSimple
   (methodRR, measure = ''mean'')
R> agg_q90<-AggregateSimple(methodRR,
   measure=''quantile'', q=0.9)
R> agg_MC <- AggregateMC(methodRR,
   type = ''MCT'', maxrank = 100)
```

While the first command merges the five considered rankings (fold-change, ordinary *t*, Limma, Fox-Dimmic *t* and shrinkage *t*), the three next lines yield aggregated rankings using the averaging procedure, the 90% quantile and the Markov chain model, respectively. Each function call produces a further ranking, which can again be analyzed with the `toplist` function.

```
R> toptenagg <- cbind(mean = toplist
   (agg_mean, show = F) $index,
q90 = toplist(agg_q90, show = F)
   $index,
MC = toplist(agg_MC, show = F)$index)
R> print(toptenagg)
```

	mean	q90	MC
[1,]	8399	8399	8399
[2,]	8225	8225	8225
[3,]	3268	8172	3268
[4,]	8172	3268	8172
[5,]	9478	8321	9478
[6,]	7106	11834	8173
[7,]	8173	7106	7106
[8,]	8321	5064	8321
[9,]	5064	9478	5064
[10,]	11834	8173	11834

As becomes obvious from this simple example, aggregation does not solve the problem of the multiplicity of possible rankings, since different aggregation schemes produce different rankings. Hence, aggregation is rather a means of averaging out the peculiarities of each single ranking than a solution to the confusing multiplicity of ranking methods. It can also be seen from the aggregation outputs that the sophisticated Markov chain method and the simple averaging procedure produce very similar results. Note that many other aggregation methods are conceivable to 'summarize' a collection of rankings. In particular, this topic is linked to spectral analysis and dimension reduction, with techniques like principal components analysis which could be used to compress several rankings into a single one. The aggregation of ranked lists is also tightly related to the voting theory, in particular the single-winner election method commonly denoted as 'Borda count' [54], which has inspired many modern aggregation procedures.

Remarks on aggregation

This article focuses on the aggregation of ranked lists. Of course, it is also possible to aggregate the ranking criterion itself and then rank the genes based on this aggregated criterion. For example, a weighted sum of the *t*-statistics obtained from subsamples can be derived [55]. A related approach ranks genes by averaging the *p*-values obtained in a large number of bootstrap samples [4]. Note that simply averaging *p*-values gives virtually more weight to the perturbed data sets for which the considered gene is not among the top-genes. Moreover, we point out that one should be cautious while interpreting *p*-values obtained from bootstrap samples, since they are known to be in average smaller than those from the original sample [56, 57]. Hence, averaged bootstrapped *p*-values should be interpreted as ranking criterion only rather than as *p*-values within a significance testing framework.

Furthermore, we note that simply averaging *p*-values, statistics or ranks obtained from different lists may miss interesting information contained in the distribution of the statistic of a particular variable across the different lists. For instance, in the case of lists obtained from *B* subsamples, not only the average statistic or rank reveal interesting patterns, but also its tails, i.e. its extreme values obtained for particular subsamples/bootstrap samples. Extreme values may indicate the presence of outliers whose removal from the data set yield completely different ranks or statistics. In this spirit, Pepe *et al.* [22] point out in the context of their 'probability of selection' framework that the whole survivor function gives a more full description of sampling variability in the ranking.

Lastly, one should be careful while interpreting aggregated ranked lists if the top-genes are highly correlated. In subsamples or bootstrap samples, variables that are highly correlated are likely to have

similar ranks. If the top-genes are highly correlated, they have to 'share' the first places and may appear at the top of the list less often than uncorrelated genes with the same level of association.

CONCLUSION

The stability of gene rankings should be routinely investigated as an important part of univariate analyses, since gene lists usually show considerable variability. In particular, it makes sense to examine the consistency of the results across different ranking criteria or different slightly modified versions of the available data set. Many procedures have been proposed in the literature for this purpose. The choice of the stability measure may depend on the seeked objective. Sophisticated approaches like the overlap score [41, 42] should probably be preferred to *ad hoc* set-based criteria if the goal is to compare the stability of two ranking criteria or the stability with respect to two different response variables. A further advantage of this approach is that it provides a solution to the difficult question of the depth of the comparison. However, simple measures such as the percentage of overlap might be more helpful to get an easily interpretable insight into the stability of the results. Aggregation, in particular aggregation of results obtained from slightly perturbed versions of the data set, is expected to yield more accurate rankings. Moreover, it is useful in practice to synthesize the results of different methods for interpretation purposes. Note, however, that there are again multiple possible aggregation schemes and thus potentially multiple aggregated rankings, as observed with the cross-entropy method using Kendall's tau and Spearman's footrule as distance measures [46]. Hence, aggregation does not intrinsically solve the problem that one does not know 'which ranking is the right one'.

Although the present article focuses on rankings, similar problems occur in the context of multivariate analyses and variable selection [38, 39, 58–60]. Before applying a 'gene signature' in clinical practice, the stability of this signature may be assessed based on resampling techniques [61]. Similarly, one can also assess the stability of, e.g. inferred network structures or clustering outputs. We actually believe that stability issues will gain much attention in future biomedical research, because they are tightly connected to the problem of 'fishing for significance' and 'false research findings' [18] in several respects. First, if the found gene list is instable with respect to the data set (i.e. if slightly modified versions of the data set yield fully different rankings), it is intuitively expected to validate poorly on a new independent data set. In this sense, the assessment of stability can be considered as a kind of preliminary (non-expensive) pseudo-validation step. Second, instability with respect to the method of analysis (i.e. when several procedures are applied successively on the same data set) increases the temptation to 'fish for significance' by trying numerous methods until one of them returns satisfying results. This approach obviously yields an optimistic bias and potentially leads to false research findings. Hence, we believe that variability across methods should be adequately acknowledged and reported in publications.

Key Points

- A ranked gene list should not be considered as a unique definitive result. It makes sense to study the stability of a list by considering alternative ranking criteria and/or slightly modified versions of the data set.
- Many stability measures and aggregation methods have been proposed to handle multiple ranked lists. Methods based on the concept of overlap and frequencies of selection are the easiest to interpret.
- Many important methods for stability assessment and aggregation of gene lists are implemented in the Bioconductor package `GeneSelector`.

Acknowledgements

We thank Martin Daumer for interesting thought-provoking discussions and the two anonymous referees for very constructive comments and suggestions that helped us to improve the manuscript.

FUNDING

The Porticus Foundation in the context of the International School of Technical Medicine and Clinical Bioinformatics. LMU-innovativ Project BioMed-S: Analysis and Modelling of Complex Systems in Biology and Medicine.

References

1. Kooperberg C, Aragaki A, Strand AD, *et al*. Significance testing for small microarray experiments. *Stat Med* 2005; **24**:2281–98.

2. Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* 2006;**7**:359.
3. Opgen-Rhein R, Strimmer K. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat Appl Genet Mol Biol* 2007;**6**:9.
4. Mukherjee S, Sykacek P, Roberts SJ, *et al*. Gene ranking using bootstrapped p-values. *ACM SIGKDD Explor Newsl* 2003;**5**:14–20.
5. Aerts S, Lambrechts D, Maity S, *et al*. Gene priorization through genomic data fusion. *Nat Biotechnol* 2006;**24**:537–44.
6. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Stat Sci* 2003;**18**:71–103.
7. Pounds S. Estimation and control of multiple testing error rates for microarray studies. *Brief Bioinform* 2006;**7**:25–36.
8. Strimmer K. A unified approach to false discovery rate estimation. *BMC Bioinformatics* 2008;**9**:303.
9. Chiaretti S, Li X, Gentleman R, *et al*. Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood* 2004;**103**:2771–8.
10. Li X. ALL. *Bioconductor R package version 1.4.4:*, 2008. http://www.bioconductor.org/packages/release/data/experiment/html/ALL.html (24 June 2009, date last accessed).
11. Slawski M, Boulesteix AL. Geneselector. *Bioconductor*, 2008. R package version 2.0.3: http://www.bioconductor.org/packages/devel/bioc/html/GeneSelector.html (24 June 2009, date last accessed).
12. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci* 2001;**98**:5116–21.
13. Efron B, Tibshirani R, Storey JD, *et al*. Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* 2001;**96**:1151–60.
14. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* 2001;**17**:509–19.
15. Lönnstedt I, Speed T. Replicated microarray data. *Stat Sin* 2002;**12**:31–46.
16. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004;**3**:3.
17. Fox RJ, Dimmic MW. A two sample Bayesian *t*-test for microarray data. *BMC Bioinformatics* 2006;**7**:126.
18. Ioannidis JPA. Why most published research findings are false. *PLoS Med* 2005;**2**:e124.
19. Boulesteix A-L, Strobl C. Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. *Technical Report 58, Department of Statistics, University of Munich*, 2009. http://epub.ub.uni-muenchen.de/10606/ (24 June 2009, date last accessed).
20. Davison AC, Hinkley DV. *Bootstrap Methods and their Applications*. New York: Cambridge University Press, 1997.
21. Politis DN, Romano JP, Wolf M. *Subsampling*. New York: Springer, 1999.
22. Pepe MS, Longton G, Anderson GL, *et al*. Selecting differentially expressed genes from microarray experiments. *Biometrics* 2003;**59**:133–42.
23. Pavlidis P, Li Q, Noble WS. The effect of replication on gene expression microarray experiments. *Bioinformatics* 2003;**19**:1620–7.
24. Xu R, Li X. A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data. *Bioinformatics* 2003;**19**:1284–9.
25. Comander J, Natarajan S, Gimbrone MA, *et al*. Improving the statistical detection of regulated genes from microarray data using intensity-based variance estimation. *BMC Genomics* 2005;**5**:17.
26. Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* 2006;**7**:55–65.
27. Tang ZQ, Han LY, Lin HH, *et al*. Derivation of stable microarray cancer-differentiating signatures using consensus scoring of multiple random sampling and gene-ranking consistency evaluation. *Cancer Res* 2006;**67**:9996–10003.
28. Qiu X, Xiao Y, Gordon A, *et al*. Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics* 2006;**7**:50.
29. Jurman G, Merler S, Barla A, *et al*. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics* 2008;**24**:258–64.
30. Gabrielsson BG, Olofsson LE, Sjögren A, *et al*. Evaluation of reference genes for studies of gene expression in human adipose tissue. *Obes Res* 2005;**13**:649–52.
31. Wilson CS, Davidson GS, Martin SB, *et al*. Gene expression profiling of adult acute myeloid leukemia identifies novel biologic clusters for risk classification and outcome prediction. *Blood* 2006;**108**:685–96.
32. Hong F, Breitling R, McEntee CW, *et al*. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* 2006;**22**:2825–7.
33. DeConde RP, Hawley S, Falcon S, *et al*. Combining results of microarray experiments: a rank aggregation approach. *Stat Appl Genet Mol Biol* 2006;**5**:15.
34. Zintzaras E, Ioannidis JPA. Meta-analysis for ranked discovery datasets: theoretical framework and empirical demonstration for microarrays. *Comput Biol Chem* 2008;**32**:38–46.
35. Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics* 2008;**24**:374–82.
36. Stolovitzky G. Gene selection in microarray data: the elephant, the blind men and our algorithms. *Curr Opin Struct Biol* 2003;**13**:370–6.
37. Zhang M, Zhang L, Zou J, *et al*. Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics* 2009;25, doi:10.1093/bioinformatics/btp295 [Epub ahead of print 5 May 2009].
38. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci* 2006;**103**:5923–8.

39. Zucknick M, Richardson S, Stronach EA. Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. *Stat Appl Genet Mol Biol* 2008;**8**:7.
40. Irizarry RA, Warren D, Spencer F, *et al*. Multiple-laboratory comparison of microarray platforms. *Nat Methods* 2005;**2**: 345–50.
41. Yang X, Bentink S, Scheid S, *et al*. Similarities of ordered gene lists. *J Bioinform Comput Biol* 2006;**4**:693–708.
42. Lottaz C, Yang X, Scheid S, *et al*. OrderedList – a bioconductor package for detecting similarity in ordered gene lists. *Bioinformatics* 2006;**22**:2315–16.
43. Fury W, Batliwalla F, Gregersen PK, *et al*. Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. In: *28th Annual International Conference of the IEEE*. USA: IEEE Engineering in Medicine and Biology Society, 2007, 5531–34.
44. Breitling R, Amtmann A, Herzyk P. Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics* 2004;**5**:34.
45. Eden E, Lipson D, Yogev S, Yakhini Z. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol* 2007;**3**:e39.
46. Lin S, Ding J. Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA studies. *Biometrics* 2009;**65**:9–18.
47. Critchlow DE. *Metric Methods for Analyzing Partially Ranked Data*. New York: Springer, 1985.
48. Zhang M, Yao C, Guo Z, *et al*. Apparently low reproducibility of true differential expression discoveries in microarray studies. *Bioinformatics* 2008;**24**:2057–63.
49. Mukherjee S, Roberts SJ, van der Laan MJ. Data-adaptive test statistics for microarray data. *Bioinformatics* 2005;**21**: ii108–14.
50. Elo LL, Filen S, Lahesmaa R, *et al*. Reproducibility-optimized test statistic for ranking genes in microarray studies. *IEEE/ACM Trans Comput Biol Bioinform* 2008;**5**: 423–31.
51. Boulesteix A-L, Strobl C, Augustin T, *et al*. Evaluating microarray-based classifiers: an overview. *Cancer Inform* 2008;**6**:77–97.
52. Meinshausen N, Bühlmann P. Stability Selection, 2008. http://arxiv.org/abs/0809.2932 (24 June 2009, date last accessed).
53. Ma S. Empirical study of supervised gene screening. *BMC Bioinformatics* 2006;**7**:537.
54. de Borda M. Mémoires sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 1781.
55. Kutalik Z, Inwald J, Gordon SV, *et al*. Advanced significance analysis of microarray data based on weighted resampling: a comparative study and application to gene deletions in Mycobacterium bovis. *Bioinformatics* 2004;**20**: 357–63.
56. Bickel PJ, Ren JJ. The bootstrap in hypothesis testing. In: *State of the Art in Probability and Statistics, Festschrift for Willem R. van Zwet, IMS Lecture Notes Monograph Series*, vol. 36. Beachwood, OH, USA: Inst of Mathematical Statistics, 2001, 91–112.
57. Strobl C, Boulesteix AL, Zeileis A, *et al*. Bias in random forest importance measures: illustration, sources and a solution. *BMC Bioinformatics* 2007;**8**:25.
58. Ein-Dor L, Kela I, Getz G, *et al*. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 2005;**21**:171–8.
59. Barla A, Jurman G, Riccadonna S, *et al*. Machine learning methods for predictive proteomics. *Brief Bioinform* 2008;**9**: 119–28.
60. Baek S, Tsai CA, Chen JJ. Development of biomarker classifiers from high-dimensional data. *Brief Bioinform* 2009;doi:10:10.1093/bib/bbp016 [Epub ahead of print 3 April 2009].
61. Davis CA, Gerick F, Hintermair V, *et al*. Reliable gene signatures for microarray classication: assessment of stability and performance. *Bioinformatics* 2006;**22**:2356–63.