Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments ☆

Rainer Breitling[a,b,*], Patrick Armengaud[a], Anna Amtmann[a], Pawel Herzyk[b,c]

[a] *Plant Science Group, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK*
[b] *Bioinformatics Research Centre, Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK*
[c] *Sir Henry Wellcome Functional Genomics Facility, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK*

Received 28 April 2004; revised 20 July 2004; accepted 22 July 2004

Available online 11 August 2004

Edited by Takashi Gojobori

Abstract One of the main objectives in the analysis of microarray experiments is the identification of genes that are differentially expressed under two experimental conditions. This task is complicated by the noisiness of the data and the large number of genes that are examined simultaneously. Here, we present a novel technique for identifying differentially expressed genes that does not originate from a sophisticated statistical model but rather from an analysis of biological reasoning. The new technique, which is based on calculating rank products (RP) from replicate experiments, is fast and simple. At the same time, it provides a straightforward and statistically stringent way to determine the significance level for each gene and allows for the flexible control of the false-detection rate and familywise error rate in the multiple testing situation of a microarray experiment. We use the RP technique on three biological data sets and show that in each case it performs more reliably and consistently than the non-parametric *t*-test variant implemented in Tusher et al.'s significance analysis of microarrays (SAM). We also show that the RP results are reliable in highly noisy data. An analysis of the physiological function of the identified genes indicates that the RP approach is powerful for identifying biologically relevant expression changes. In addition, using RP can lead to a sharp reduction in the number of replicate experiments needed to obtain reproducible results.

Keywords: Microarray analysis; Differentially expressed gene

1. Introduction

DNA microarrays are a powerful technology for monitoring the expression level of thousands of genes simultaneously. They provide the basis for a variety of applications, including tumor classification, molecular pathway modeling, and functional genomics. One of the most popular uses, however, is the comparison of gene expression difference under two distinct experimental conditions (treated vs. untreated samples, diseased vs. normal tissue, mutant vs. wild-type organisms, etc.). In this kind of experimental setup, the main challenge is the identification of those genes whose expression is significantly different between the two conditions. The earliest approach used a simple fold-change (FC) criterion to detect genes of interest [1], but this has the obvious disadvantage that it does not provide a significance estimate for the observed changes and that the necessary cutoff values (2-fold? 1.68-fold?) are essentially arbitrary. Hence, replicated experiments and increasingly sophisticated statistical tests were soon suggested to achieve a more reliable identification of differentially regulated genes. While all of these techniques provide results that are better justified than a crude FC heuristics, they have several shortcomings. First of all, they lack the intuitive appeal of the FC criterion. In many situations, it is unlikely that very small fold-changes have any biological relevance even if they are significant statistically, for example when many other genes in the same condition show much larger changes. Second, the authors of the different algorithms offer little, if any, justification for their underlying error models. Indeed, modern techniques which make relatively weak assumptions give more robust results than a classical *t*-test that assumes normality of the error, but they still in most cases assume a symmetrically distributed error [2]. This assumption is clearly violated by the biological variation in a number of the most common experimental designs, e.g., in a tumor-versus-normal comparison every tumor sample may contain some amount of normal tissue, so that the real value of interest, i.e., the expression in a pure tumor sample, lies well outside the measured values. The third shortcoming of current statistical techniques is that the reasons for their differences in performance are very little understood. Most current approaches use similar basic statistics and differ mainly in their determination of the significance/rejection level [2]. When applied to biological data (instead of simulated data sets), they give very similar overall results [2–4], but often show important and seemingly random differences for some genes [2–4]. There is currently no convincing rationale for choosing between the different approaches. The problem is further aggravated by the statistical expertise necessary to appreciate the differences between the various algorithms. This has led to a widespread feeling among biologists that the choice of statistical analysis is rather arbitrary and

☆ Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2004.07.055.

* Corresponding author. Fax: +44-141-330-4447.
E-mail addresses: r.breitling@bio.gla.ac.uk (R. Breitling), p.armengaud@bio.gla.ac.uk (P. Armengaud), a.amtmann@bio.gla.ac.uk (A. Amtmann), p.herzyk@bio.gla.ac.uk (P. Herzyk).

Abbreviations: RP, rank product; SAM, significance analysis of microarrays; FC, fold-change; HDL, high-density lipoprotein

0014-5793/$22.00 © 2004 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2004.07.055

that the resulting gene lists are far from reliable. These are serious concerns that have discouraged some biologists from considering microarray analysis for their particular scientific questions.

Here, we present a novel statistical technique derived from biological reasoning to overcome some of these problems. Our technique, which is explained in detail below, is based on the calculation of rank products (RP), is statistically rigorous and can be used to provide reliable significance thresholds to distinguish significantly regulated genes. The results and simplicity of RP are surprisingly similar to fold-changes, but overcome its most significant limitations. In contrast to previous approaches, the assumptions made for our method are relatively weak. We assume (1) that relevant expression changes affect only a minority of genes, (2) measurements are independent between replicate arrays, (3) most changes are independent of each other, and (4) measurement variance is about equal for all genes. The latter is a biologically reasonable assumption that can be met by a number of recent normalization techniques [5–7], which are rapidly becoming a standard in the field, independent of the analysis method used. Assumption (1) may look relatively restrictive, as it is easy to imagine experimental setups where a majority of genes are differentially expressed. However, even in experiments where very many genes are changed, one will usually want to know the most important ones; so what seems to be a limitation is indeed a very useful feature. In addition, it is a conservative assumption as it tends to restrict the number of genes that will be called significantly changed.

We compare the performance of our method on three biological data sets with the significance analysis of microarrays, SAM [8]. We chose SAM as our benchmark, because of its reported performance in comparative studies and because it is a prototypical example of a class of current techniques with similar concept [2,9,10]. According to the Science Citation Index, SAM is currently the most popular method for differential gene expression analysis. Here, we show that our method yields more sensitive and reliable results than SAM when evaluated by a variety of criteria.

2. Materials and methods

2.1. Analytical techniques

The RP method. Our algorithm for identifying differentially expressed genes is based on the approach that a biologist might make when analyzing the data manually: Consider a simple two-color microarray experiment comparing mRNA levels under conditions A and B on one slide. After the first experiment, the biologist will know which genes are most highly up- or downregulated in condition A vs. B, but, as the data are very noisy, one would not rely too much on these results. However, if the same gene shows up at the top of the list in a replicate experiment, one's confidence will increase and after consistent results in further replicates may reach virtual certainty. For an experiment examining n genes in k replicates, one might argue that the probability for a certain gene to be at the top of each list (rank 1) is exactly $1/n^k$ if the lists were entirely random. Put differently, it is extremely unlikely to observe a single gene at the top position of all replicates if none of the genes were differentially expressed, i.e., if all null hypotheses were true. More generally, for each gene g in k replicates i, each examining n_i genes, one can calculate the corresponding combined probability as a rank product $\mathrm{RP}_g^{\mathrm{up}} = \prod_{i=1}^{k}(r_{i,g}^{\mathrm{up}}/n_i)$, where $r_{i,g}^{\mathrm{up}}$ is the position of gene g in the list of genes in the ith replicate sorted by decreasing FC, i.e. $r^{\mathrm{up}} = 1$ for the most strongly upregulated gene, etc. Analogously, $\mathrm{RP}_g^{\mathrm{down}}$ is calculated from the list of genes sorted by increasing FC, i.e. $r^{\mathrm{down}} = 1$ for the most strongly downregulated gene. Note that when $n_i = n$ for all replicates, one can alternatively use the geometric mean rank $\bar{r}_g^{\mathrm{up}} = (\prod_{i=1}^{k} r_{i,g}^{\mathrm{up}})^{1/k}$, without losing any information. These RP values (or $\bar{r}$ values) can then be used to sort the genes according to the likelihood of observing them so high on the lists of differentially expressed genes just by chance. In many cases, this will already be sufficient: Genes with the smallest RP values are the most interesting candidates and the biologist can then select some of them for further study. This method is sufficiently straightforward to be implemented manually in Excel for a typical number of replicates (see Supplementary Material).

For single-channel arrays, e.g., Affymetrix GeneChip® arrays, the RP values are calculated over all possible pairwise comparisons. To correct for the fact that the pairwise comparisons between samples are not independent, the significance analysis has to be adjusted as described below.

Determination of significance levels. Usually, it will be interesting to know how significant the changes are and how many of the selected genes are likely to be truly differentially expressed. Consider a microarray experiment with two replicates (A and B), each examining n genes. In that case, the RP for a certain gene g will be $\mathrm{RP}_g = (\mathrm{rank}_g^{\text{replicate A}}/n) \times (\mathrm{rank}_g^{\text{replicate B}}/n)$. This can be interpreted as a p-value, as it describes the probability of observing gene g at a certain rank ($\mathrm{rank}_g^{\text{replicate A}}$) or better in the first replicate and at another rank ($\mathrm{rank}_g^{\text{replicate B}}$) or better in the second replicate. Note that this interpretation is valid when all ranks are equally likely, which is the case when the replicates are independent, genes have equal variance and none of them are differentially expressed. These are exactly the assumptions of our method as described. The reason for why this p-value (= RP value) cannot be used directly to assess the significance of an observed expression change is simple: We are not interested in the probability that a gene shows a certain expression pattern, but in the combined probability, p', of *all* expression patterns that are as unlikely. In simple cases, i.e., when the number of genes is small, this probability can be directly calculated from the RP values. For example, in an experiment with two replicates and three genes, if a gene g is observed at rank 2 in replicate A and in rank 1 in replicate B, its RP value will be $\mathrm{RP} = (2/3) \times (1/3) = 2/9$. The same RP value can be obtained if the gene has rank 1 in the first and rank 2 in the second replicate. Thus, the combined probability, which we are interested in, is $p' = 2\mathrm{RP} = 4/9$. In general, for k replicates and n genes, this probability can be calculated by multiplying the RP value by a factor F, where F is the number of possible products of k numbers smaller than n that are equal to the numerator of the RP value. In the above example, F equals 2 because there are two such products (1×2 and 2×1). F depends on the number and type of prime factors of the numerator of the RP value, thus it is not easy to calculate in general.

Fortunately, a simple permutation-based estimation procedure provides a very convenient way to determine how likely it is to observe a given RP value or better in a random experiment, thus converting from the RP value to an E value in analogy to the BLAST results familiar to molecular biologists [11]. In the test cases described below, we approximate the RP value distribution in each case by calculating the RP values for, say, 100 random "experiments" with the same number of replicates and "genes" as the real experiment. Each random experiment consists of k random permutations of the numbers $1, \ldots, n$ and for these the RP values are calculated as described above. We can then just count how many simulated RP values smaller than or equal to a given experimental RP value occur in the 100 random experiments ($x(\mathrm{RP})$) and calculate the average expected value $E(\mathrm{RP}) \approx x(\mathrm{RP})/100$. Large numbers of random experiments will provide better estimates of the E value. At the same time, the procedure to convert RP values to E values automatically addresses the multiple testing problem associated with the simultaneous analysis of thousands of genes in the same experiment.

Subsequently, for each gene g, one can also calculate a conservative estimate of the percentage of false-positives (PFP) if this gene (and all genes with RP values smaller than this cutoff) would be considered as significantly differentially expressed: $q_g = E(\mathrm{RP}_g)/\mathrm{rank}(g)$. Here, rank(g) denotes the position of gene g in a list of all genes sorted by increasing RP value, i.e., it is the number of genes accepted as significantly regulated. This estimates the false discovery rate [12] and provides a flexible way to assign a significance level to each gene. One can now decide how large a PFP would be acceptable and extend the list of accepted genes up to the gene with this q_g value.

The approach described above, which corresponds to determining the false discovery rate in SAM, assumes the independence of genes and may yield underestimates when expression measurements of many of the top genes on an array are dependent, e.g., because of cross-hybridization. However, this will not affect the sorting of genes but only the decision of where to place a significance cutoff.

For single-channel arrays, the random experiment has to be slightly modified. Instead of permuting the ranks for each replicate, we perform permutations of the expression values for each array and then calculate the corresponding fold-changes for all pairwise comparisons. This replicates the dependency structure of the original data. These fold-changes are then used to calculate RP values as described above.

Normalization and pre-processing. For the analysis by SAM and by the RP method, exactly the same normalized data were used. For consistency, all data (see below) were subjected to quantile normalization [13]. To normalize the measurement variance at different intensities, we used variants of the started-log procedure [5], i.e., for the *A. thaliana* data we used the intensity values directly, without background subtraction, for the leukemia data, we added a constant to all measurements, so that the smallest value becomes one, and for the ApoAI/SR-BI data we added a constant of 7680, which we estimated to minimize the deviation from constant variance using the method outlined in [5]. We intentionally used these simplified techniques to avoid introducing particular assumptions in the normalization process that would bias the performance of the detection algorithms in the next step. For the same reason, no filtering for low-intensity or low-quality signals was done.

SAM. The implementation of the SAM algorithm [8] was obtained as a Microsoft Excel add-in from http://www-stat.stanford.edu/~tibs/SAM/.

2.2. Data sets

Arabidopsis thaliana potassium starvation. This data set from our laboratory examines plant seedlings that were germinated and grown in the absence of the essential macro-nutrient potassium for 2 weeks versus non-starved control samples of the same age and tissue. The experiment consists of 3 two-color microarrays for shoots and 3 two-color microarrays for roots. Each array contains 1250 non-blank probes for genes encoding known and predicted transmembrane transporters, as well as controls, all spotted *in duplicate* in two subarrays as described in [14]. A detailed biological interpretation of these data will be published elsewhere.

High-density lipoprotein (HDL)-deficient mouse models. This study examined gene expression in two mouse models with low levels of HDL. Expression data for ApoAI knockout and SR-BI transgenic mice [15,16] were obtained from http://www.stat.berkeley.edu/users/terry/zarray/Html/matt.html. Each data set contains eight hybridizations of Cy5 labeled mRNA from wild-type mouse liver and eight hybridizations of Cy5 labeled mutant mouse mRNA, each hybridized against a pool of Cy3 labeled mRNA from the same eight wild-type mice. A total of 6384 genes and controls were present on each array.

Acute leukemia samples. Expression data for bone-marrow samples from leukemia patients [17] were obtained from http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi. The data set consists of 27 hybridizations of acute lymphoblastic leukemia (ALL) samples and 11 hybridizations of acute myeloid leukemia (AML) samples on Affymetrix high-density oligonucleotide microarrays representing 7129 genes and controls.

3. Results

The comparison of the performance of algorithms for the detection of differentially expressed genes is hampered by the difficulty of deciding on a "gold standard" experiment. For a biological sample it is usually not known which genes are the true positives, and simulated data or arrays that are "spiked" with known amounts of certain mRNAs are both highly artificial and not necessarily representative of the conditions encountered in real samples. Even the "verification" of microarray results by an alternative technique such as quantitative RT-PCR is just replacing one error-prone method by another. One might define a true positive gene as one that is highly likely to be detected as differentially expressed each time the same experiment is repeated, and we make use of this concept below. In addition, we use biological knowledge to assess the quality and consistency of the results. As a benchmark for our RP approach, we used the SAM [8], one representative of a class of current analytical techniques that have all been shown to have quite similar performance [2,4]. As in most analyses below the predicted number of differentially expressed genes is small, SAM should perform particularly well [2], In some cases, we also included the results of an average-FC ranking of genes, which despite its obvious shortcomings concerning the definition of significant changes performs quite well for many biological data sets.

3.1. Arabidopsis thaliana *potassium starvation*

In the first attempt to evaluate the consistency of our algorithm, we made use of our *A. thaliana* membrane transporter microarray [14], which is spotted in two identical subarrays and thus contains an internal duplicate of each gene. We analyzed the data for each duplicate spot independently, i.e., we treated the two subarrays as if they contained different genes. Of course the measurements for the duplicate spots are highly dependent, but we argue that if a true positive is a gene that has an increased likelihood of being detected as differentially expressed in any repetition of the experiment, then it should be even more likely to be detected in two simultaneous subarrays. A good algorithm would be one that consistently detects differentially expressed genes in both duplicate spots.

The results are shown in Table 1. It can be seen that, overall, SAM and RP detect about the same number of

Table 1
Comparison of the number of "spots" identified as differentially expressed in potassium-starved *A. thaliana* seedlings

	Total spots identified at FDR 10%	Number of genes	Duplicated detections[a]
SAM			
Shoots	0	0	0
Roots	68	55	24 (35%)
Roots plus shoots	65	53	22 (34%)
RP			
Shoots	32	21	22 (69%)
Roots	36	24	24 (67%)
Roots plus shoots	78	52	49 (63%)
	Total spots included[b]		
Fold-change			
Shoots	32	22	20 (63%)
Roots	36	22	27 (75%)
Roots plus shoots	78	54	53 (68%)

The number of spots that correspond to genes found in duplicate is also shown. See text for details about the combined roots plus shoots analysis. FDR false discovery rate.

[a] A few genes are represented by three or more spots.

[b] For the FC method, the number of genes included was chosen to be the same as for RP. This corresponds to minimum fold-changes between 1.5-fold (roots + shoots, upregulated) and 2.1-fold (roots, upregulated).

differentially expressed spots at a predicted false-discovery rate of 10%. The performance of RP is slightly more even, in that it detects differentially expressed genes in shoots (where SAM finds no significant differences between the samples at this cutoff level) and also detects some downregulated genes. The main difference is in the ''duplicate recovery rate'', i.e., the percentage of spots that are also identified as regulated in their corresponding duplicate in the second subarray. In SAM this rate is about 1/3, which means that about 20% of the detected genes are found in duplicate. The corresponding rate in RP is about twice as high, about 2/3, and a simple pen-and-paper calculation shows that this means that about half of the genes are found in duplicate. This result does not change if it is corrected for the different number of genes detected as ''significantly changed'' by the two methods. It can also be seen that RP and FC perform very similarly – except that FC is naturally not usable to delimit the length of the results lists based on a significance criterion. Instead, the cutoff from RP had to be used for this purpose to make the comparison possible.

Fig. 1 shows a more detailed comparison of the results of SAM, RP and a FC ranking. For the latter, the cutoff was selected in such a way that the resulting gene lists had the same length as for RP. The same was done for SAM in the case of shoots, as SAM did not detect significantly changed genes at a false discovery rate of 10% in that tissue. It can be seen that the majority of genes that are detected by all three methods are found in duplicate by at least one of them. This seems to confirm the idea that these genes are enriched for true positives. It is also striking that there is a large overlap between RP and FC, which again contains a large number of genes detected in duplicate. This good agreement was not necessarily expected, as the RP method requires a radical transformation of the raw data and discards all information about the absolute expression levels. It shows, however, that RP yields results that agree well with the intuitive concept that interesting genes will show the highest consistent changes.

On the other hand, the overlap between RP and SAM is larger than between FC and SAM, which may be due to a higher ''robustness'' of RP towards inconsistent changes, such as single outliers. Some limits of the performance parameters of SAM and RP may be estimated from the observed ''duplicate recovery rate''. If for the moment we assume that all spots called ''significantly different'' in duplicate correspond to true positives and those that are detected only once are either all true positives (and should be detected twice) or all false positives (and should not be detected at all), we can estimate limits for the PFP and the percentage of false-negatives (PFN) in these two extreme scenarios. In the case of SAM (duplicate recovery rate about 1/3), either the PFN $\geqslant$ 40% (and the PFP $\geqslant$ 0%) or the PFP $\geqslant$ 66% (and the PFN $\geqslant$ 0%). For RP (where the duplicate recovery rate is about 2/3), the corresponding limit values are PFN $\geqslant$ 25% (PFP $\geqslant$ 0%) and PFP $\geqslant$ 33% (PFN $\geqslant$ 0%), respectively. The true values are probably somewhere in-between, but in any case RP seems to perform better than SAM in this respect.

As a special challenge for the statistical tests, we not only analyzed roots and shoots separately but also included a joined analysis (roots plus shoots), which should detect genes that are consistently changed in both tissues. In the joined data, the measurement variance for each gene is

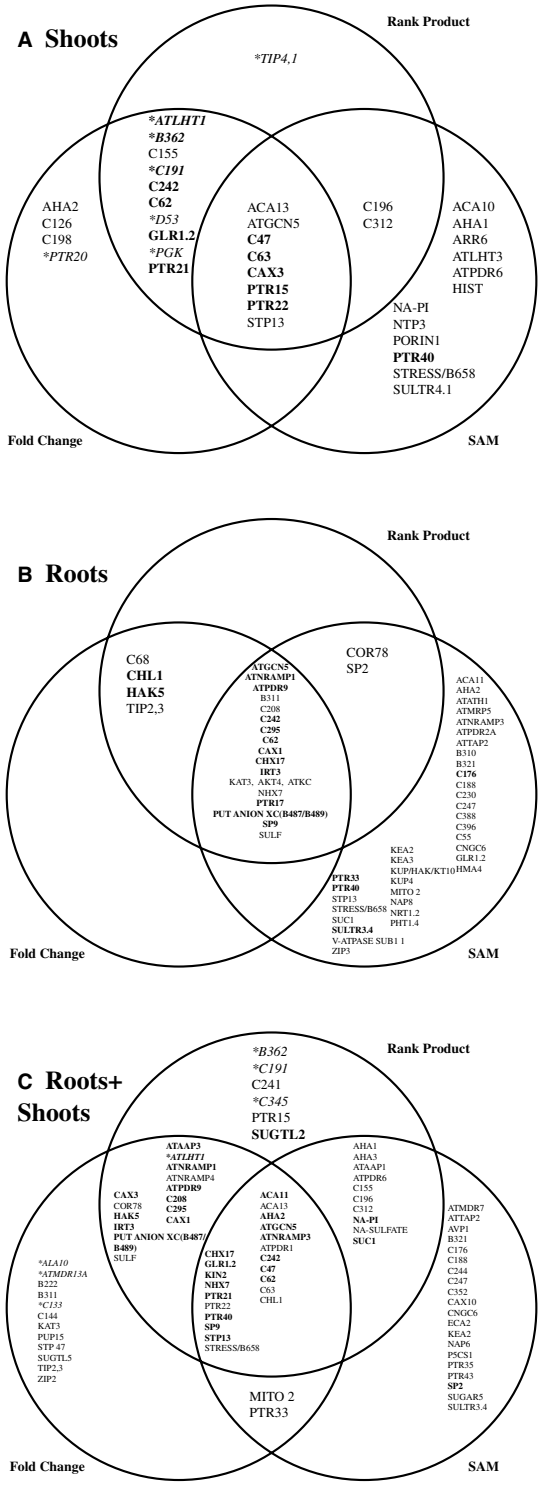

Fig. 1. Venn diagrams showing the genes identified as differentially regulated by the three algorithms indicated, in potassium-starved *A. thaliana* shoots (A), roots (B) and an analysis of both data sets combined (C, see text for details). Genes that were identified in duplicate by at least one method are shown in bold. Genes that are downregulated in the potassium-starved condition are marked by *. Where available, genes are indicated by their gene names, predicted and uncharacterized genes by arbitrary identifiers. Although SAM did not detect significantly changed genes in shoots, the top 22 genes found by SAM are included for comparison.

Table 2
Comparison of the top 20 differentially expressed genes in HDL-deficient mouse models identified by RP and SAM

ID	Rank product method	ID	SAM method
(A) *Top 20 up-regulated genes in ApoAI k.o. (not significant at a false detection rate of ⩽20%)*			
1238§	Similar to cytochrome *c* oxidase polypeptide I	1239§	Similar to *S*-adenosylmethionine synthetase
1239§	Similar to *S*-adenosylmethionine synthetase	5692	Olf-1
2939†	Apolipoprotein A-II precursor	4418§	Similar to fructose-bisphosphate aldolase B
4346	Hemoglobin β, pseudogene	1254§	Est
4418§	Similar to fructose-bisphosphate aldolase B (human)	1238§	Similar to cytochrome *c* oxidase polypeptide I
1248		417	Similar to cAMP-dependent protein kinase major membrane substrate
3488	Procollagen, type IV, α1	6029	
847		2929†	Apolipoprotein C-I precursor
1222	Est	5540	Est
3386	MDB0087	4892	FGF12A
5986	Cy3RT	1235§	Similar to cytochrome *b*
6136†	Peroxisomal fatty acyl CoA oxidase	5043	Blank
809	Est	45	
1254§	Est	5100	Est
1242		3445	cDNA from an individual with schizophrenia
799	Cy3RT	2813	Est
1235§	Similar to cytochrome *b*	5918	Ezrin
6008		862§	Similar to cytochrome *c* oxidase polypeptide I
839	Similar to ATP synthase A chain	840	
862§	Similar to cytochrome *c* oxidase polypeptide I	5539	Est
(B) *Top 20 down-regulated genes in ApoAI k.o.*			
2149*	Apo AI	2537*	Highly similar to apolipoprotein C-III precursor
540*	Highly similar to apolipoprotein A-I precursor	2149*	Apo AI
5356*	Catechol *O*-methyltransferase	4139*	Weakly similar to C-5 sterol desaturase
2537*	Highly similar to apolipoprotein C-III precursor	5356*	Catechol *O*-methyltransferase
4139*	Weakly similar to C-5 sterol desaturase	1739*	Apo CIII
1739*	Apo CIII	1496*	Est
1496*	Est	540*	Highly similar to apolipoprotein A-I precursor
563†	Fatty acid-binding protein	4941*	Similar to yeast sterol desaturase
3729†	Moderately similar to fatty acid-binding protein	1652	Est
1337†	Psoriasis-associated fatty acid binding protein	2107§	Sox5
2091	??	2106§	Pigment-epiththelium derived factor
2106§	Pigment-epiththelium derived factor	2083	Unk homeo3
2538†	Apolipoprotein B mRNA-editing component 1 (Apobec1)	1347	Transcription factor GATA-6
2323†	Colony stimulating factor, macrophage	2165	Tumor necrosis factor
1736†	30 kDa adipocyte complement-related protein Acrp30	2536†	Retinol-binding protein II, cellular
4941*	Similar to yeast sterol desaturase	2022	Similar to sarcoplasmic reticulum histidine-rich calcium-binding
3709	ID-2 DNA binding protein inhibitor/extra macrochaetae homolog1	2048	
3346†	Hormone-sensitive lipase	3623	
2107§	Sox5	3729†	Moderately similar to fatty acid-binding protein
4530	Brain protein D3	2160†	CTP: phosphocholine cytidylyl transferase
(C) *Top 20 up-regulated genes in SR-BI transgenic*			
1581*	SRB1 mouse PCR	1581*	SRB1 mouse PCR
783*	SRB1 mouse PCR	783*	SRB1 mouse PCR
3261*	Similar to glutathione *S*-transferase π class	3261*	Similar to glutathione *S*-transferase π class
5312§	Diff. Ass. Prot. 13 kDa	5312§	Diff. Ass. Prot. 13 kDa
2287§	Ubiquitin-conjugating enzyme E2H	3615§	
2408*	Similar to hemoglobin ζ chain (human)	3709§	ID-2 DNA binding protein inhibitor/extra macrochaetae homolog1
3*	mSRB1	976	MDB0008
4921*	Scavenger receptor class B type I (mSR-BI)	4966§	MDB0211
4006*	Similar to hemoglobin α chain	4003*	Similar to hemoglobin β-1 chain
4815*	Similar to hemoglobin β chain	2408*	Similar to hemoglobin ζ chain
3709§	ID-2 DNA binding protein inhibitor/extra macrochaetae homolog1	5649	Similar to cytochrome *c* oxidase polypeptide II
421§	Est	4006*	Similar to hemoglobin α chain
3615§		4815*	Similar to hemoglobin β chain
5203*	β globin	936†	Highly similar to fatty acid-binding protein
409	Est	421§	Est
4013		4407	Est
471*	Similar to hemoglobin α chain	2459*	Similar to α-globin
4087	Unk	2287§	Ubiquitin-conjugating enzyme E2H

Table 2 (*continued*)

ID	Rank product method	ID	SAM method
2459*	Similar to α-globin	5923	Even skipped homeotic gene 2 homolog
4966§	MDB0211	1293	IGF2R
(D) *Top 20 down-regulated genes in SR-BI transgenic*			
4285§	Capping protein α2	4755§	Angiotensinogen
2854	Similar to human metallothionein-Ie gene	4629	MDB1406
4755§	Angiotensinogen	5759†	Long chain fatty acid CoA ligase
4866	est	2482§	Myogenin-like
5676	tyk2 like	6135†	Plasminogen activator inhibitor-2, macrophage
5243	est	2523†	Macrophage colony stimulating factor I receptor precursor
5914	est	4137	Highly similar to tubulin γ-1 chain
1267†	Similar to very low density lipoprotein receptor	1267†	Similar to very low density lipoprotein receptor
4483	Unk homeo	2873§	Ex312 (CTG4a)
2401		3120	Retinoblastoma 1
2482§	Myogenin-like	4285§	Capping protein α2
2873§	Ex312 (CTG4a)	3134	Est
4527†	Weakly similar to fibrinogen α and α-E chain precursors	2543†	TGF β receptor-II
5759	Long chain fatty acid CoA ligase	3154	Est
1075§	Growth factor receptor bound protein 2	2856	Est
5262	Est	1075§	Growth factor receptor bound protein 2
5325†	Retinol binding protein (RBP)	4133	Neurofilament, heavy polypeptide
4934†	HMG CoA synthase	1745	Highly similar to cytochrome *c* oxidase polypeptide VA precursor
2127	Highly similar to cytochrome P450 IIC24	4482	Apterous homolog 2
867†	Similar to diazepam-binding inhibitor	1206	

(E)

Summary	RP	SAM	Fold-change
Blanks/ Controls	10	8	6
Genes identified by *t*-test (*)	19	16	20
Additional lipid-related genes (†)	15	10	16

Genes reported as differentially expressed by Dudoit et al. [15] based on a strict *t*-test are marked by a *. Additional lipid-related genes are marked by a †, further genes shared by SAM and RP are marked by a §. Table 2E shows a summary, comparing the number of blanks/controls identified by each method (probable false positives), the number of genes also identified by the *t*-statistics (probable true positives) and further lipid-related genes (possible true positives). The ID numbers are arbitrary identifiers.

expected to be increased, thus making the detection of significant changes more difficult. In fact, using the criteria outlined above the advantages of RP over both SAM and FC are particularly obvious in this analysis (Fig. 1C), probably because the relative fold-changes represented by ranks are more robust than the absolute fold-changes considered by SAM and FC.

3.2. HDL-deficient mouse models

In the second analysis, we compare the performance of SAM and RP on expression data from mouse models with low levels of HDL. In this case, the underlying biology of the experiment is well understood and can be used to assess the quality of the algorithms. Due to the more complex experimental design of that study [16], which uses a common reference instead of paired hybridizations between wild-type and mutant samples, the RP method had to be slightly modified. We calculated the RP values independently for the comparison of wild type vs. reference pool (RP_{wt}) and for mutant vs. reference pool (RP_{mut}) and then used the ratio RP_{mut}/RP_{wt} to sort the genes, thus eliminating genes that are changed to the same extent and direction in both mutant and wild-type samples relative to the reference pool. Note that as always this procedure is performed independently for up- and downregulation. The results are shown in Table 2.

All methods agree that there are no significantly upregulated genes in the ApoAI knockout, confirming that the *E* value as calculated for RP performs comparable to the SAM (the top 20 genes that are shown in Table 2A were not significant at a FDR $\leqslant$ 20%). Both SAM and RP find the same genes that have already been reported as changed by a strict *t*-test procedure [15], RP finds slightly more replicate probes for these genes among its list of highest scoring genes. More significant, however, is the detection rate of lipid-related genes among the top 20 hits for the two methods, as the biology of the mouse models indicates that such genes are most likely to be true positives [16]. RP includes about 50% more genes of this class among its top 20 hits. This is particularly striking in the case of the ApoAI knockout, where three replicate spots for fatty-acid binding protein (FABP) receive very low (i.e., significant) RP values, while only one of them is included among the top 20 hits reported by SAM. The fact that three replicate spots for FABP are assigned consistently low RP values indicates that this gene is indeed a true positive and should not be missed by a reliable detection algorithm. The same is probably true for

Table 3
Comparison of the performance of RP and SAM on subsets of the leukemia data set of Golub et al. [17]

	SAM (all data)	RP (all data)	RP1	RP2	RP3	SAM1	SAM2	SAM3
Top 25 up-regulated genes								
Zyxin	1	1	15	37	15	56	15	70
FAH fumarylacetoacetate	2	3	88	61	104	153	1	75
Leukotriene C4 synthase (LTC4S) gene	3	12	17	67	45	2	84	15
LYN V-yes-1 Yamaguchi sarcoma viral-related oncogene homolog	4	4	43	1691	39	107	1555	38
CTSD cathepsin D (lysosomal aspartyl protease)	5	7	2	92	14	29	383	236
FTL ferritin, light polypeptide	6	19	4997	3	4	1107	17	154
APLP2 amyloid β (A4) precursor-like protein 2	7	11	115	157	32	414	692	132
DF D component of complement (adipsin)	8	13	566	658	47	426	1067	1
Induced myeloid leukemia cell differentiation protein MCL1	9	21	959	234	20	691	357	48
CST3 cystatin C (amyloid angiopathy and cerebral hemorrhage)	10	2	26	507	1	971	1256	31
PRG1 proteoglycan 1, secretory granule	11	59	1	5495	6	14	2241	5
LEPR leptin receptor	12	5	51	272	175	143	1126	430
CD33 CD33 antigen (differentiation antigen)	13	14	42	2050	66	53	2288	20
PPGB protective protein for β-galactosidase (galactosialidosis)	14	29	13	2555	41	1	1941	114
Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA	15	67	90	9	1496	298	16	732
ATP6C vacuolar H+ ATPase proton channel subunit	16	23	852	23	30	693	69	50
Interleukin-8 precursor	17	49	3057	1	12	832	38	34
Interleukin 8 (IL8) gene	18	44	3015	4	35	802	22	121
ITGAX integrin, α X (antigen CD11C (p150), α polypeptide)	19	16	36	136	1749	4	652	1595
GB DEF = homeodomain protein HoxA9 mRNA	20	17	1154	119	93	1320	35	120
ME491 gene extracted from *H. sapiens* gene for Me491/CD63 antigen	21	6	60	2839	28	225	1647	216
Liver mRNA for interferon-γ inducing factor (IGIF)	22	73	199	1389	241	39	1580	26
P58 natural killer cell receptor precursor mRNA, clone cl-39	23	20	14	4711	11	3549	3853	1595
PFC properdin *P* factor, complement	24	45	37	5060	78	135	3068	18
Peptidyl-prolyl *cis–trans* isomerase, mitochondrial precursor	25	42	638	219	181	890	175	361
Top 25 downregulated genes								
ALDR1 aldehyde reductase 1 (low K_m aldose reductase)	1	10	81	147	120	154	282	342
Retinoblastoma binding protein P48	2	6	36	58	37	30	70	499
Macmarcks	3	2	10	152	6	72	623	1
C-myb gene extracted from human (c-myb) gene, complete primary cds, and five complete Alternatively spliced cds	4	1	452	5	2	445	63	85
ACTN2 actinin α2	5	71	611	1631	203	864	1779	621
Oncoprotein 18 (Op18) gene	6	42	620	11	48	691	94	344
IEFSSP 9502 mRNA	7	34	681	244	192	776	217	289
HNRPG heterogeneous nuclear ribonucleoprotein G	8	31	92	1076	104	103	1091	837
Proteasome ι Chain	9	8	11	3	20	276	4	427
MYL1 myosin light chain (alkali)	10	24	128	256	750	10	183	1641
GCND3 cyclin D3	11	4	8	10	1	47	121	15
ACADM acyl-coenzyme A dehydrogenase, C-4 to C-12 straight chain	12	12	70	134	53	76	343	143
Inducible protein mRNA	13	9	44	126	737	214	790	446
Stimulator of TAR RNA binding (SRB) mRNA	14	11	132	155	140	285	602	677
Estrogen sulfotransferase mRNA	15	139	2120	2556	12	3076	229	136
Ras GTPase-activating-like protein (IQGAP1) mRNA	16	37	570	42	113	665	108	316
Pulmonary surfactant-associated protein a precursor	17	5	195	61	32	3076	2531	274
HMG1 high-mobility group (non-histone chromosomal) protein 1	18	148	4745	691	8	1237	804	199
Transcription factor (CBFB) mRNA, 3′ end	19	47	96	103	1768	196	262	1451
SRP9 signal recognition particle 9 kDa protein	20	22	32	105	40	39	289	1567
MCM3 minichromosome maintenance deficient (*S. cerevisiae*) 3	21	27	131	487	282	194	707	1497

Table 3 (*continued*)

	SAM (all data)	RP (all data)	RP1	RP2	RP3	SAM1	SAM2	SAM3
Heat shock protein, 70 kDa (Gb:Y00371)	22	39	55	89	3171	8	429	2959
Interferon γ up-regulated 1-5111 protein precursor	23	50	25	26	2437	27	86	3539
DHPS deoxyhypusine synthase	24	53	583	196	429	577	85	645
Butyrophilin (BTF5) mRNA	25	29	166	113	410	234	120	146
Average rank	13	24.08	555.62	737.32	313.56	536.52	721.8	506.66
Median rank	13	20.5	91	141.5	47.5	219.5	316	207.5
			Median rank(RP): 99.5			Median rank(SAM): 235		

The top 25 up- and downregulated genes identified by SAM of the complete data set are shown. The first two columns show the ranks assigned to each of these 50 genes by SAM and RP in their analysis of the complete data set. The next columns show the ranks assigned by RP (RP1–RP3) and SAM (SAM1–SAM3) in their analysis of random subsets of the data, each consisting of three lymphoblastic and three myeloid samples. The average and median rank assigned for each subset and the median for all three subsets is indicated at the bottom of the table.

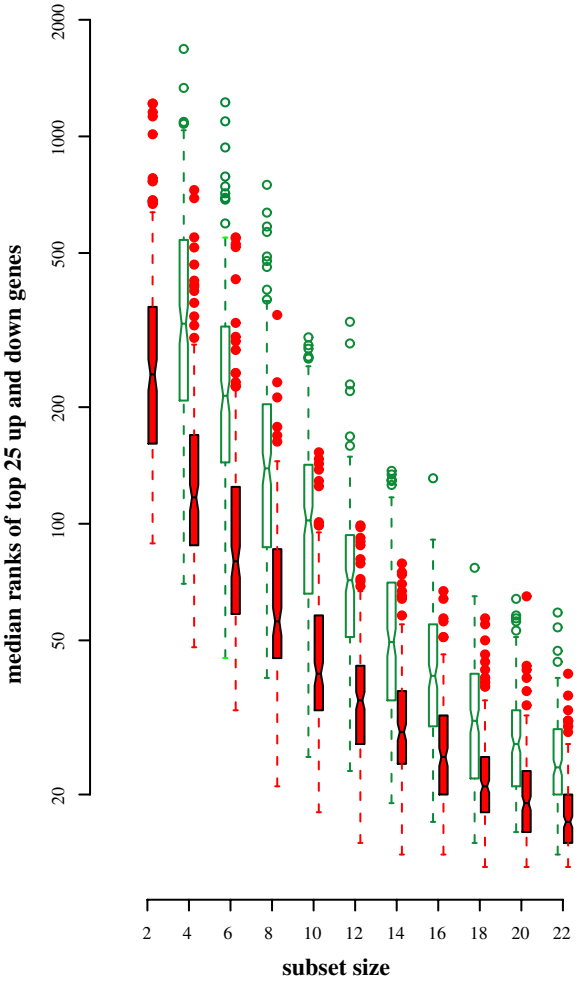

Fig. 2. Boxplot showing the dependence of RP and SAM results on the number of replicates. The boxes extend from the first to the third quartile, with a notch at the median. The "whiskers" extend to 1.5× the interquartile range but not further than the data range. If the central notches do not overlap, the medians are significantly different at the 5% level. SAM requires replication, therefore no SAM results are shown for a subset size of 2 (i.e., an unreplicated experiment). See text for details. RP filled symbols; SAM open symbols.

ApoBecl and Adiponectin, which are likely to be functionally relevant for an HDL-deficient mouse model [18,19], yet are missed by SAM and the classical *t*-test (see Table 2E for a summary).

As in the *A. thaliana* example, FC and RP show a very similar performance. This demonstrates that RP can indeed serve as a reasonable replacement of FC, while at the same time overcoming its essential limitations (such as arbitrary threshold and lack of statistical interpretability).

3.3. Acute leukemia samples

In both the *A. thaliana* data and the HDL-deficient mouse mutant data, our results might be explained by a lower false-negative rate in the RP method, which might come at the cost of a corresponding increase in the false-positive rate, although there is no indication that the false-positive rate of RP is particularly high. In the third analysis, we used an entirely different comparison to avoid this problem.

The leukemia data set [17] contains a large number of high-quality replicates. We argue, in agreement with Broberg [4], that an analysis of the complete data set should yield a reliable approximation to a "gold standard" set of true positives. A good algorithm should then give good ranks to these true positive genes even when using only a small fraction of the complete data set [4]. We thus chose three random subsets of the data, each consisting of three ALL and three AML samples and analyzed them by SAM and RP. As RP works with paired data, we used all nine possible pairwise combinations of the 3 + 3 data sets as input for that method. For each subset, we determined the rank assigned to the top 25 up- and down-regulated genes determined by a SAM analysis of the complete data set (note that SAM gets a head start in this respect). The results are shown in Table 3. It can be seen that SAM and RP agree quite well in their analysis of the complete data. However, in the analysis of the subsets, RP performs much better: The median rank assigned to the "gold standard" genes is about twice as good as that assigned by SAM. This difference is much more striking than those observed by Broberg [4] in his comparison of SAM, a Bayesian approach [20] and his *samroc* algorithm using a similar approach. This confirms once more the sensitivity and reliability of the RP method.

3.4. Dependence on the number of replicates

The large number of hybridizations in the leukemia data set makes it possible to check how SAM and RP perform depending on the number of replicates used for the analysis (Fig. 2). As before, we use the complete data set as a substitute for a "gold standard". We selected the top 25 up- and down-regulated genes by either SAM or RP of the complete data. Then, we determined the rank of these genes in random subsets of the data (each consisting of equal numbers of myeloid and lymphoblastic samples). The median rank of the 25 genes converges to 13 when an increasing fraction of the total data is used. As can be seen in Fig. 2, RP converges much faster on its final result than SAM. For example, on average a simple duplicate experiment (subset size = 4 arrays) analyzed by RP is equivalent to a quadruplicate experiment (subset size = 8 arrays) analyzed by SAM.

4. Discussion

We have presented evidence that the simple if unconventional RP method outperforms a sophisticated statistical technique (SAM) as assessed by a variety of criteria. As the assumptions of SAM and other modern techniques for the detection of differentially regulated genes are closely related [2] and their performance in comparative studies is quite similar, we predict that this finding is not restricted to SAM but is of more general relevance.

What can explain the success of RP? Several factors probably contribute. First, RP makes only relatively weak assumption about the data, i.e., it expects about equal variance for all genes. This requirement is biologically reasonable and is easily met by a number of recently developed normalization techniques [5–7], the simplest of which (started-log [5]) we used in the present study. In contrast, even the weak assumptions made by other non-parametric statistics may be too strong for microarray data. Second, our results seem to confirm the biological intuition that significant gene regulation is almost a switch-like process, so that relevant changes should always be large, while small changes may have statistical but rarely biological significance. Previously, the strong disagreement between, e.g., *t*-test results and simple FC lists, has been used as an argument for the superiority of the former. In contrast, we argue that the surprisingly good agreement between RP and the average-FC approach increases our confidence in the novel method. The difference between RP and FC criteria is nonetheless very important: not only is RP more robust against outliers, it also provides a simple and straightforward procedure to determine the significance of an observed change. The *A. thaliana* data show that the significance threshold determined in this way can correspond to widely different absolute fold-changes.

The strong performance of RP with very small data sets (Fig. 2) indicates that another advantage of RP is its stability against increasing levels of uncertainty in the variance estimation that seriously hamper all variants of the *t*-test, such as SAM. RP does not rely on estimating the measurement variance for each single gene and thus is particularly useful when this estimate becomes unreliable due to a low number of replicates. But as the analysis of the Leukemia data set shows, even for 11 replicates RP can outperform SAM significantly.

It is obvious that the RP method has other advantages over previous statistical techniques, not the least being the simplicity of its underlying reasoning. Another advantage is the ease with which results from various experimental platforms can be combined in one analysis. As long as the results are expressible as ranked lists, it does not matter whether they are produced by two-color cDNA arrays, filter arrays, or Affymetrix oligo chips. It can also be applied to proteome and metabolome studies where ranked lists of changed proteins or metabolites are produced by 2D-gels or mass spectrometry. The method also provides suitable input data for automated microarray interpretation tools such as iterative Group Analysis [21]. And of course the RP method could have wider applications as a new non-parametric test in the statistical analysis of diverse rankable data outside of biology.

Rank products provide a powerful new test statistics for defining differentially expressed genes in microarray experiments. Because of its non-parametric nature, it requires only a few well-justified assumptions about the data. In contrast to previous techniques, in particular *t*-test-based statistics such as SAM, RP does not depend on an estimate of the gene-specific measurement variance and is therefore particularly powerful when only a small number of replicates are available, as is currently the case in the majority of biological studies. In such cases, it can cut the number of required hybridizations in half (Fig. 2), thus resulting in considerable savings of both time and money (or rather expanding the range of microarray applications to areas where more replicates are not available for various reasons). Our analysis also demonstrates the usefulness of real data sets, in contrast to simulated data, for the assessment of algorithm performance in microarray analysis. The RP method has been successfully implemented and is now used as the standard method for microarray data analysis at the Sir Henry Wellcome Functional Genomics Facility at University of Glasgow (http://www.gla.ac.uk/functionalgenomics/).

Acknowledgements: We thank Dr. Daniela Höller and Dr. Ernst Wit, and members of the Amtmann laboratory for helpful discussions. This work was supported by BBSRC Grants 17/G17989 (to A.A., P.H., and Ernst Wit) and 17/P17237 (to A.A.).

References

[1] DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997) Science 278, 680–686.
[2] Pan, W. (2002) Bioinformatics 18, 546–554.
[3] Troyanskaya, O.G., Garber, M.E., Brown, P.O., Botstein, D. and Altman, R.B. (2002) Bioinformatics 18, 1454–1461.
[4] Broberg, P. (2003) Genome Biol. 4, R41.
[5] Rocke, D.M. and Durbin, B. (2003) Bioinformatics 19, 966–972.
[6] Durbin, B.P., Hardin, J.S., Hawkins, D.M. and Rocke, D.M. (2002) Bioinformatics 18 (Suppl. 1), S105–S110.
[7] Huber, W., Von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. (2002) Bioinformatics 18 (Suppl. 1), S96–S104.
[8] Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Proc. Natl. Acad. Sci. USA 98, 5116–5121.
[9] Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001) J. Am. Statist. Assoc. 96, 1151–1160.
[10] Thomas, J.G., Olson, J.M., Tapscott, S.J. and Zhao, L.P. (2001) Genome Res. 11, 1227–1236.

[11] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) J. Mol. Biol. 215, 403–410.
[12] Storey, J.D. (2003) Ann. Statist. 31, 2013–2035.
[13] Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003) Bioinformatics 19, 185–193.
[14] Maathuis, F.J. et al. (2003) Plant J. 35, 675–692.
[15] Dudoit, S., Yang, Y.H., Callow, M.J. and Speed, T.P. (2002) Statist. Sinica 12, 111–139.
[16] Callow, M.J., Dudoit, S., Gong, E.L., Speed, T.P. and Rubin, E.M. (2000) Genome Res. 10, 2022–2029.
[17] Golub, T.R. et al. (1999) Science 286, 531–537.
[18] Diez, J.J. and Iglesias, P. (2003) Eur. J. Endocrinol. 148, 293–300.
[19] Teng, B., Ishida, B., Forte, T.M., Blumenthal, S., Song, L.Z., Gotto Jr., A.M. and Chan, L. (1997) Arterioscler. Thromb. Vasc. Biol. 17, 889–897.
[20] Lonnsted, I. and Speed, T. (2002) Statist. Sinica 12, 31–46.
[21] Breitling, R., Amtmann, A. and Herzyk, P. (2004) BMC Bioinformatics 5, 34.