FEBS Letters

FEBS PRESS science publishing by scientists

# A cautionary note on the rank product statistic

James A. Koziol

Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA, USA

**Correspondence**
J. A. Koziol, Department of Molecular and Experimental Medicine, The Scripps Research Institute, MEM290, 10550 N. Torrey Pines Road, La Jolla, CA 92037, USA
Fax: +1 858 784 8658
Tel: +1 858 784 2704
E-mail: koziol@scripps.edu

(Received 7 February 2016, revised 11 April 2016, accepted 22 April 2016, available online 1 June 2016)

doi:10.1002/1873-3468.12194

Edited by Takashi Gojobori

**The rank product method introduced by Breitling R *et al.* [2004, FEBS Letters 573, 83–92] has rapidly generated popularity in practical settings, in particular, detecting differential expression of genes in microarray experiments. The purpose of this note is to point out a particular property of the rank product method, namely, its differential sensitivity to over- and underexpression. It turns out that overexpression is less likely to be detected than underexpression with the rank product statistic. We have conducted both empirical and exact power studies that demonstrate this phenomenon, and summarize these findings in this note.**

**Keywords:** one-sided tests; rank product statistic; symmetry of test statistics; two-sided tests

In an influential paper, Breitling and colleagues [1] introduced a valuable nonparametric statistical technique for detecting differentially regulated genes in replicated microarray experiments. Their rank product method entails ranking expression levels within each replicate, then computing the product of the ranks across the replicates. The rank product is then compared to its sampling distribution under a permutation model for subsequent inference. Although developed originally for microarrays analyses, the rank product method has found widespread acceptance in such diverse settings as meta-analysis and proteomics [e.g., 2,3,4], with PubMed citations in the hundreds.

## The rank product statistic

We briefly describe the rank product statistic. We start with expression levels for $n$ genes from $k$ replicate samples. The expression level for the $i$th gene in the $j$th replicate by $X_{ij}$, where $1 \le i \le n$, $1 \le j \le k$ is denoted. Next, we rank the expression levels $X_{1j}, X_{2j}, \ldots, X_{nj}$ in each replicate $j$, forming $R_{ij} = \text{rank}(X_{ij})$, $1 \le R_{ij} \le n$. [The ranking is such that genes with the smallest ranks are the 'most interesting' from a biological perspective.] Then, Breitling's rank product statistic for the $i$th gene is, up to a normalization constant, the product

$$RP_i = \prod_{j=1}^{k} R_{ij}.$$

Genes associated with sufficiently small $RP$ values would be marked for further consideration, and Breitling and colleagues [1] posit a statistical formalism for such a determination. In particular, they propose a permutation approach to calculate the distribution of the $RP_i$ under the null hypothesis that the $X_{ij}$ are identically distributed (exchangeable) within each of the $k$ independent replicates. More recently, Eisinga *et al.* [5] and Heskes *et al.* [6] have devised numerical algorithms for exact calculation of the null distribution of the rank product statistic, without reliance on permutations.

An equivalent statistic to $RP_i$ is the monotone transformation

$$\log(RP_i) = \sum_{j=1}^{k} \log(R_{ij}) = -k * \log(n+1) + \sum_{j=1}^{k} \log(R_{ij}/(n+1)) \quad (1)$$

Monotonicity ensures that achieved significance levels of $RP_i$ and $\log(RP_i)$ are identical. Note that, since the $k$ replicates are independent, $\log(RP_i)$ is the sum of $k$ independent, identically distributed random

variables under the null hypothesis, which can lead to more transparent distributional theory [7].

The term $-k$*log$(n + 1)$ in (1) is a constant, so for simplicity we focus on the statistic

$$LRP_i = -\sum_{j=1}^{k} \log(R_{ij}/(n+1)) \qquad (2)$$

in the remainder of this note, and refer to it generically as the log rank product statistic. [We have multiplied by $-1$ to ensure that $\mathrm{LRP}_i$ will be positive].

## Operating characteristics of the rank product statistic

In Breitling's original formulation, ranks near 1 (overexpression of a particular gene) were considered most interesting: ranks near 1 would lead to small values of his original rank product statistic, or correspondingly large values of the log rank product statistic (2); hence, a one-sided test statistic with rejection region based on suitably small values of the rank product statistic or large values of (2) would be appropriate.

Suppose instead that gene under expression is also of interest. In Breitling's ranking scheme, ranks near $n$ connote underexpression, and one might accordingly formulate a one-sided test statistic for underexpression with rejection region based on suitably small values of the log rank product statistic (2).

We have investigated both over- and underexpression alternatives relative to the log rank product statistic (2); details are given in the Appendix. It turns out that the log rank product statistic is differentially sensitive to these two classes of alternatives, being much more sensitive to underexpression than overexpression.

In practice, what are the ramifications? There is no conceptual difficulty in formulating two-sided tests based on the log rank product statistic (2). However, the distribution of this statistic is not symmetric, so separate consideration of the alternatives of under- and overexpression is warranted. [As a reviewer has pointed out, asymmetry in the null distribution of the log rank product statistic is especially apparent with small $k$.] We remark that, if one expects to detect any deviation in expression levels with the rank product statistic (either over- or underexpression) then one might expect the problem ought to be invariant to ranking method: that is, one could choose to rank expression levels from 1 through $n$, but with 1 designating the lowest expression, $n$ the highest, and arrive at the same conclusions as with Breitling's original rank ordering. Investigators need to be cognizant of the differential sensitivity of the rank product statistic to over- and underexpression, and perhaps adjust their test procedures to accommodate alternatives of particular interest.

## Acknowledgements

The author thanks the reviewers for their insightful comments, which led to a much improved presentation. This research was supported in part by National Institutes of Health grant 1 P01 HL119165.

## References

1 Breitling R, Armengaud P, Amtmann A and Herzyk P (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* **573**, 83–92.
2 Breitling R and Herzyk P (2005) Rank-based methods as a non-parametric alternative of the t-test for the analysis of biological microarray data. *J Bioinf Comp Biol* **3**, 1171–1189.
3 Hong F, Breitling R, McEntree CW, Wittner BS, Nemhauser JL and Chory J (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* **22**, 2825–2827.
4 Hong F and Breitling R (2008) A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics* **24**, 374–382.
5 Eisinga R, Breitling R and Hestes T (2013) The exact probability distribution of the rank product statistics for replicated experiments. *FEBS Lett* **587**, 677–682.
6 Heskes T, Eisinga R and Breitling R (2014) A fast algorithm for determining bounds and accurate approximate p-values of the rank product statistic for replicate experiments. *BMC Bioinformatics* **15**, 367.
7 Koziol JA (2010) Comments on the rank product method for analyzing replicated experiments. *FEBS Lett* **584**, 941–944.
8 Fisher RA (1925) Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.
9 Feller W (1971) An Introduction to Probability Theory and its Applications, Vol. **II**, 2nd edn. John Wiley and Sons, New York, NY.
10 Wallace DL (1959) Bounds on normal approximations to Student's and the chi-square distributions. *Ann Math Statist* **30**, 1121–1130.

## Appendix

## Simulation study

For simplicity, we focus on the expression patterns of one particular gene from a set of $n$ genes, in $k$ replicate experiments. Let $R_j$ denote the rank of this gene's

expression in the $j$th experiment, $1 \le j \le k$. Under the null hypothesis of no rank association across the $k$ independent samples, the distribution of each $R_j$ will be uniform over the integers $\{1, 2,\ldots, n\}$, hence will have the expected value $(n + 1)/2$. If this gene is overexpressed relative to the remaining $n-1$ genes, each $R_j$ would likely be nearer 1 than expected; and if this gene is underexpressed, each $R_j$ would likely be nearer $n$ than expected.

For power considerations, we posit a distribution for the $R_j$ under patterns of over- or underexpression, derived from the beta distribution. The probability density function of the beta distribution $B(a,b)$ is given by

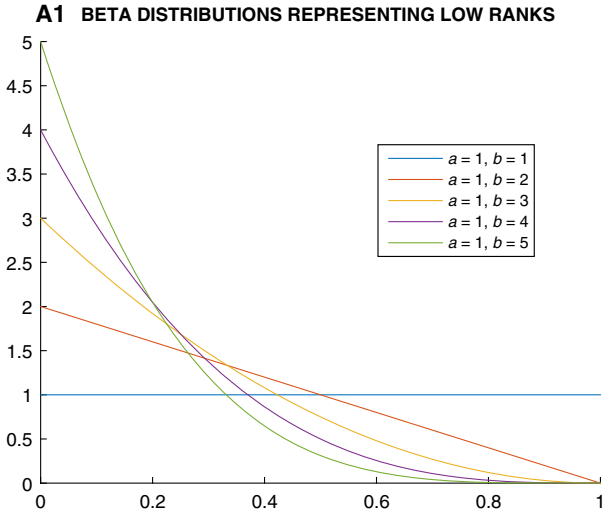
$$f(x;a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \qquad \text{(A1)}$$

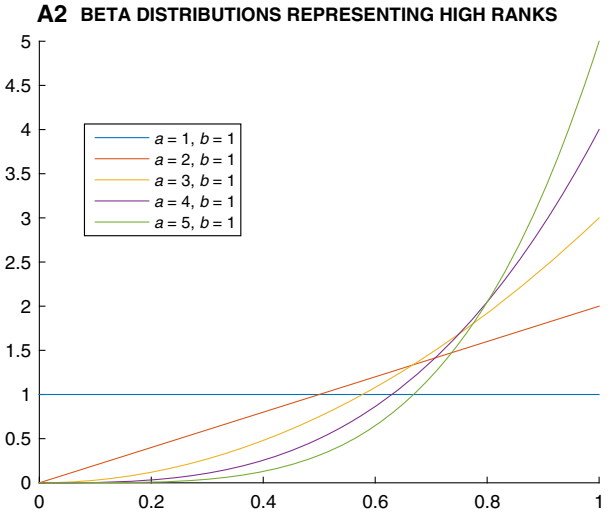
over the unit interval $0 \le x \le 1$, where $G$ denotes the gamma function, and $a > 0$ and $b > 0$ are shape parameters. The beta distribution is continuous, whereas the distribution of ranks in any experiment is discrete over the set $\{1, 2,\ldots, n\}$, hence some explanation is warranted. Starting with a $B(a,b)$ distribution over the unit interval [0, 1], the unit interval is divided into $n$ equal subintervals with cutpoints $1/n$, $2/n,\ldots$, 1; then, $\Pr(R_j) = i$ is taken as the $B(a,b)$ probability of the $i$th subinterval, $1 \le i \le n$. We remark that the uniform distribution of ranks over $\{1, 2,\ldots, n\}$ is thereby generated from the $B(1,1)$ distribution (uniform distribution over [0, 1]).

Besides uniformity of ranks with $a = 1$, $b = 1$, ranks tending to be nearer 1 would be represented by beta distributions with $a = 1$, $b > 1$, and ranks nearer the maximum would be represented by beta distributions with $a > 1$, $b = 1$. We depict these prototypical beta distributions in Figs A1 and A2, respectively.

With $a = 1$, as the shape parameter $b$ increases, the ranks will increasingly tend to be nearer 1 (overexpression); and with $b = 1$, as the shape parameter $a$ increases, the ranks will increasingly tend to be nearer the maximum $n$ (underexpression). Figs A1 and A2 are reflections of each other about the line $x = 1/2$, as might be anticipated from inspection of equation A1: the probability density function exhibits reflection symmetry through the relation $f(x,a,b) = f(1-x,b,a)$. Hence, there is equivalence of representations of overexpression in ranks with $B(1,c)$ distributions and representations of underexpression in ranks with $B(c,1)$ distributions, for $c > 1$.

The simulated power curves for the rank product statistic under alternatives representing over- or underexpression (Fig A3 and A4) represent averages of 10 000 runs, for each choice of $k$, and $n = 1000$. For computational purposes, we draw random variables from the prescribed $B(a,b)$ distributions, and use a ceiling function on the scaled random variates to simulate the ranks. Also, we use the log rank product statistic (2) and calculate exceedances relative to one-sided 0.05-level critical values from the gamma $G(k,1)$ approximation to its null distribution [7]. All simulations were undertaken in Matlab R2014b (Mathworks Inc., Natick, MA, USA). Overexpression corresponds to smaller ranks and larger values of the LRP statistic, and underexpression corresponds to




**Figs. A1 and A2.** Beta $B(a,b)$ probability density functions (equation A1) over the unit interval [0,1] for selected values of $(a,b)$. The $B(a,b)$ distributions are used to generate ranks $R_i$ in replicate experiments. The $B(1,b)$ distributions (Fig. A1) will generate ranks near 1, representing overexpression of gene i, whereas the $B(a,1)$ distributions (Fig. A2) will generate ranks near $n$, representing underexpression of gene i. $B(a,b)$ and $B(b,a)$ probability density functions are reflections of each other about $x = 1/2$.

larger ranks and smaller values of the LRP statistic. As expected, power increases with $k$ (the number of replicate experiments), and as the beta distributions increasingly deviate from the null $B(1,1)$ distribution. This is easily seen in Figs A3 and A4. But closer comparison of Figs A3 and A4 reveals a salient fact: power of the log rank product statistic is substantially higher against $B(c,1)$ distributions than for $B(1, c)$ distributions; for the normalized ranks: the log rank product statistic is more sensitive to underexpression than to overexpression.

## Theoretical considerations

We provide some theoretical underpinnings for the empirical power study we have described previously. We assume $n$ is sufficiently large that the normalized ranks $R_j/(n+1)$ are (approximately) distributed as $B(a,b)$. Under the null hypothesis that the $R_j$ are equidistributed over $\{1, 2,\ldots, n\}$, the normalized ranks are (approximately) uniformly distributed over the unit interval [0,1], corresponding to the $B(1,1)$ distribution. Hence, the log transformation $-\log(R_j/(n+1))$ has an (approximate) exponential distribution, so the log rank product statistic for the particular gene under consideration

$$LRP = \sum_{j=1}^{k} -\log(R_j/(n+1)) \qquad \text{(A2)}$$

has a gamma distribution, $G(k,1)$ [7]. We remark that we could have taken $-2*\log(R_j/(n+1))$ instead, in analogy with Fisher's method for combining $P$-values [8]: then, twice the log rank product statistic (A2) would have an approximate $X^2_{2k}$ null distribution. In fact, Fisher's combination procedure provides clear motivation for Breitling's original formulation of the rank product statistic, once analogy is drawn between ranks and '$P$-values' small ranks (near 1) induce small $P$-values (near 0) through normalization by sample size $n$, engendering large values for Fisher's combination procedure.

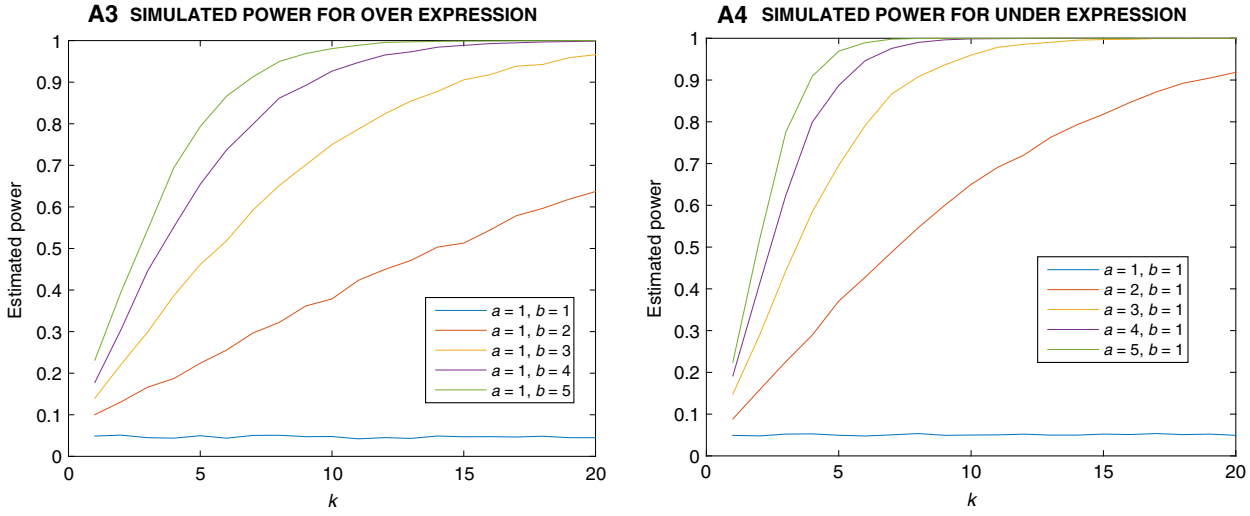
We consider the two classes of alternative distributions for the normalized ranks:

(i) $a = 1, b > 1$ (overexpression, Fig. A1)

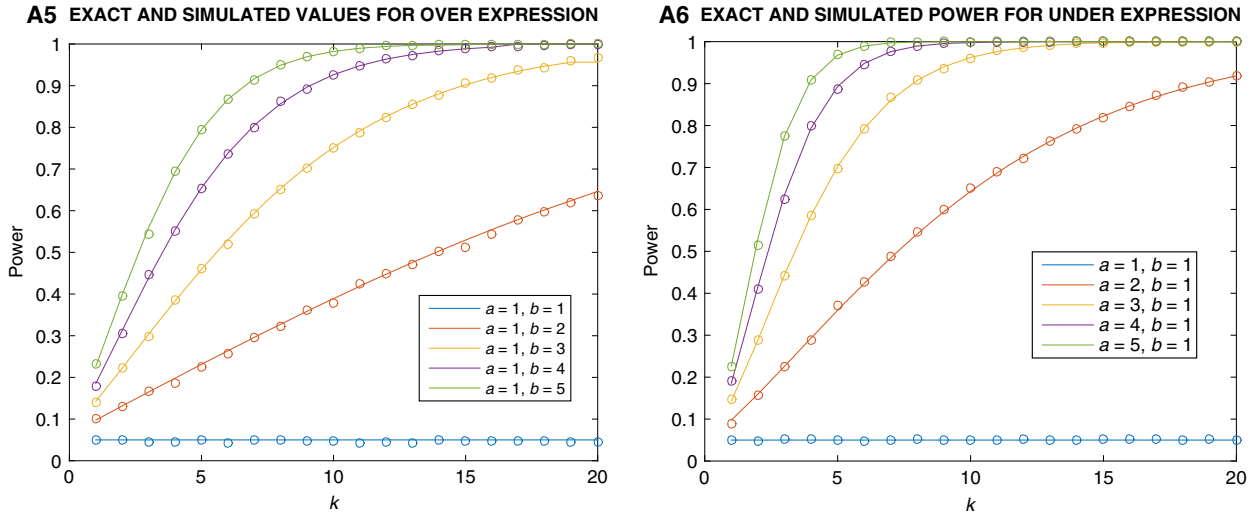
(ii) $a > 1, b = 1$ (underexpression, Fig. A2)

These are power law distributions: under (i), the normalized ranks would tend to be nearer 0 than a uniform distribution, and under (ii), the normalized ranks would be nearer 1 than a uniform distribution.

Let us first consider the alternative (ii), for which exact distribution theory is available. Under a $B(a,1)$




**Figs. A3 and A4.** Simulated power of the log rank product statistic (A2) with beta distributions generating the ranks. The power curves represent averages of 10 000 runs, for each choice of $k$ (number of replicate experiments) and $n = 1000$ (number of genes). The designated beta distribution $B(a,b)$ was used to generate the ranks $R_i$ prior to calculation of the log rank statistic (equation A2). The estimated power values are the exceedances of the log rank statistic relative to one-sided 0.05-level critical values from the gamma $G(k,1)$ approximation to its null distribution [7]. Simulations from $B(1,1)$ correspond to the null distribution of the log rank product statistic, hence exceedances here should be close to the nominal 0.05 level. For any $B(a,b)$ alternative, power increases as $k$ (the number of replicate experiments) increases. For fixed $k$, power increases as the $B(a,b)$ alternative becomes more 'extreme'. Power against any $B(c,1)$ alternative (underexpression) exceeds the power against the analogous $B(1,c)$ alternative (overexpression).




**Figs. A5 and A6.** Exact and simulated power of the log rank product statistic, for varying $k$, and selected $B(a,b)$ alternatives. The simulated power values (open circles) are taken from the simulation study (Figs A3 and A4). The exact power values are derived either from the exact distribution of the log rank product statistic against $B(a,1)$ alternatives (Fig. A6, underexpression) or by numerical integration of the inverse Laplace transform of the log rank product statistic against $B(1,b)$ alternatives (Fig. A5, overexpression). With few exceptions as noted in the text, exact and simulated values are within 0.01 of each other.

distribution, each $-\log(R_j/(n+1))$ will have an exponential distribution with shape parameter $(1 + a)$, hence the log rank product statistic log $RP_k$ from equation (A2) has a gamma distribution, $G(k,1 + a)$. Using this representation, we can compute the power of the log rank product statistic exactly; we used Matlab R2014b for these calculations. In Fig. A6 we plot the simulated and exact powers of the log $RP_k$ statistic under the alternatives depicted in Fig. A2. Agreement is excellent, with only two points $((k,a) = (5,2)$ and $(2,5))$ having differences greater than 0.01 (but less than 0.015) in absolute magnitude between the simulated and exact values.

Unfortunately, distribution theory attending alternative (i) is much more involved, since closed form distributional representations are rather complex. Instead, we will rely on Laplace transforms for power calculations of the log rank product statistic under alternative (i). (Laplace transforms are closely related to moment generating functions, but are particularly useful with continuous random variables having positive support. Feller [9] gives a cogent overview of Laplace transforms in probability theory.) We outline this approach as follows.

Suppose $X$ has a $B(1,b)$ distribution. Then the Laplace transform $L(t;b)$ of $X$ is given by

$$L(t;b) = \frac{\Gamma(b+1)\Gamma(1+t)}{\Gamma(1+t+b)}$$

Note that $L(t;b)$ simplifies further for $b$ an integer. It follows that the Laplace transform of the log $RP_k$ statistic is $L(t;b)^k$, so the probability density function of the log $RP_k$ statistic is the inverse Laplace transform of $L(t;b)^k$. Numerically, the inverse Laplace transform can be calculated in Matlab or Mathematica, and the exact power of the log $RP_k$ statistic can then be determined by numerical integration of the probability density function over the appropriate critical region (which depends on $k$). We used Mathematica 10 (Wolfram Research Inc., Champaign, IL, USA) for these calculations. In Fig. A5 we plot the simulated and exact powers of the log $RP_k$ statistic under the alternatives depicted in Fig. A2. As with the underexpression alternatives, agreement is excellent, with only two points $((k,b) = (4,2)$ and $(6,3))$ having differences greater than 0.01 (but less than 0.015) in absolute magnitude between the simulated and exact values.

## Asymptotic distribution theory

In typical genomic and proteomic applications, there are generally a large number of genes ($n$), but few replicates ($k$). As a reviewer has presciently pointed out, in such settings the distribution of the log rank product statistic under the null hypothesis of equidistribution of ranks is asymmetric, leading to the differential power properties documented in this note. But the null distribution of the log rank product statistic is

asymptotically normal and therefore approaches symmetry with increasing $k$, rendering moot the issue of differential sensitivity to various alternatives. The asymptotic normality follows directly from the central limit theorem; or, using Fisher's transform to a chi-squared distribution, from the convergence of the chi-squared distribution to normality as degrees of freedom increase. This in turn leads to useful bounds on the goodness of the normal approximation, for example, [10].

## An example

Let us conclude with a simple example1 that should illustrate the main message of this note. Take a common scenario, assessing $n = 1000$ genes with $k = 5$ replicates. We suspect gene A is overexpressed, and find its ranks across the replicates to be 100, 150, 200, 250, and 300. For this gene, the log rank statistic (2) turns out to be 8.40, with corresponding one-sided (right tail) $P$-value 0.079. We might thereby conclude that there is little evidence to support overexpression of gene A. In comparison, we also suspect gene B to be underexpressed, and observe its ranks across the replicates to be 900, 850, 800, 750, and 700. In other words, the magnitude of departures from an 'expected' rank of 500.5 under a null hypothesis of equidistribution of ranks is the same for gene A and gene B. But for gene B, the log rank statistic (2) is 1.14, with corresponding one-sided (left tail) $P$-value of 0.006. In short, the evidence for underexpression of gene B is much stronger than the evidence of overexpression of gene A, even though the ranks are mere reflections of one another.