Statistical Applications in Genetics and Molecular Biology

Volume 5, *Issue* 1 2006 *Article* 15

Combining Results of Microarray Experiments: A Rank Aggregation Approach

Robert P. DeConde[*] Sarah Hawley[†] Seth Falcon[‡]
Nigel Clegg[**] Beatrice Knudsen[††] Ruth Etzioni[‡‡]

[*] Public Health Science Division, Fred Hutchinson Cancer Research Center, Seattle, WA, rdeconde@gmail.com

[†] Public Health Science Division, Fred Hutchinson Cancer Research Center, Seattle, WA, shawley@fhcrc.org

[‡] Program in Computational Biology, Fred Hutchinson Cancer Research Center, Seattle, WA, sfalcon@fhcrc.org

[**] Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA, ncleg@fhcrc.org

[††] Fred Hutchinson Cancer Research Center, bknudsen@fhcrc.org

[‡‡] Public Health Science Division, Fred Hutchinson Cancer Research Center, Seattle, WA, retzioni@fhcrc.org

Combining Results of Microarray Experiments: A Rank Aggregation Approach[*]

Robert P. DeConde, Sarah Hawley, Seth Falcon, Nigel Clegg, Beatrice Knudsen, and Ruth Etzioni

Abstract

As technology for microarray analysis becomes widespread, it is becoming increasingly important to be able to compare and combine the results of experiments that explore the same scientific question. In this article, we present a rank-aggregation approach for combining results from several microarray studies. The motivation for this approach is twofold; first, the final results of microarray studies are typically expressed as lists of genes, rank-ordered by a measure of the strength of evidence that they are functionally involved in the disease process, and second, using the information on this rank-ordered metric means that we do not have to concern ourselves with data on the actual expression levels, which may not be comparable across experiments. Our approach draws on methods for combining top-*k* lists from the computer science literature on meta-search. The meta-search problem shares several important features with that of combining microarray experiments, including the fact that there are typically few lists with many elements and the elements may not be common to all lists. We implement two meta-search algorithms, which use a Markov chain framework to convert pairwise preferences between list elements into a stationary distribution that represents an aggregate ranking (Dwork et al, 2001). We explore the behavior of the algorithms in hypothetical examples and a simulated dataset and compare their performance with that of an algorithm based on the order-statistics model of Thurstone (Thurstone, 1927). We apply all three algorithms to aggregate the results of five microarray studies of prostate cancer.

KEYWORDS: rank aggregation, microarrays, meta-analysis, Markov chains, order-statistic models

[*] Research supported by P50 CA 97186 from the National Cancer Institute. Corresponding author: Ruth Etzioni, Fred Hutchinson Cancer Research Center, Mailstop M2-B230, 1100 Fairview Ave North, Seattle, WA 98109-1024. retzioni@fhcrc.org

1 Introduction

Widespread use of high-throughput genomic and protein analysis is providing researchers with the opportunity to combine (or aggregate) results across sets of experiments designed to explore the same biological phenomenon. These sets of experiments may consist, for example, of microarray studies conducted by different groups that have compared gene expression patterns under similar sets of conditions (Rhodes et al, 2002, 2004) (e.g. cancerous versus normal tissue or progressive versus non-progressive tumors). They may attempt to aggregate measures of expression at the DNA, RNA, or protein levels (Varambally et al, 2005). And they may even attempt to aggregrate results across diseases or species (McCarroll et al, 2004).

Combining information across multiple studies is challenging for many reasons. In the case of microarray studies, the use of different technologies means that not all studies measure expression levels of the same genes. In addition, technical, biological, and other sources of variability will generally lead to measurements of gene expression that are not comparable across studies.

To avoid dealing directly with measurements of gene expression that may not be comparable, several approaches have been proposed. Rhodes et al (2002, 2004) computed q-values (Benjamini and Hochberg, 1995) for each gene, and defined a differential expression signature for each of S experiments as the set of genes with q-values below a pre-defined threshold. The meta-signature was declared to be all genes present in at least J signatures, where J was selected by permutation testing. Applying this approach to analyze microarry results from 36 studies comparing cancerous versus normal tissue, Rhodes et al found 183 genes present in at least 10 signatures, while under their permutation null distribution zero such genes were expected.

A second approach (Wang et al, 2004) used differences between the log gene expression ratios under each experimental condition. These differences were computed for each gene and each experiment and combined across experiments in a Bayesian fashion. Final inferences were based on standard hypothesis testing procedures applied to the combined differences. A third approach used information on the correlation between gene expression measurements (Parmigiani et al, 2004). Rather than providing an aggregate inference, this approach focused on identifying a set of comparable genes, namely genes for which the correlation (of expression values) with other genes in the array was similar across studies.

In this article, we base analysis on the ranked list of genes produced by each study. The rankings reflect the results of statistical hypothesis tests for differential expression across the experimental conditions of interest and represent an ordering of the genes in terms of priority for further study. Our use of ranked lists represents another attempt to move away from gene expression measurements to a metric that we anticipate will be more comparable across studies. Indeed, a recent study (Yuen et al, 2002) compared microarray measurements between Affymetrix©GeneChips and two-color cDNA microarrays and found that, although the fold changes of differentially expressed genes showed poor correlation across array platforms, the rank orders of differentially expressed genes were comparable.

In the ranked-list metric, meta-analysis corresponds to aggregating the rankings across studies. From a statistical perspective, we may think of attempting to estimate a modal, or central, ranking that summarizes the distribution of the observed rankings of the genes across studies. There is an extensive and mature literature on statistical models for ranked data, which includes the order-statistic models of Thurstone (Thurstone, 1927, 1931), paired-comparison models (Smith, 1950; Bradley and Terry, 1952), and multi-stage models (Plackett, 1975; Fligner and Verducci, 1988). Generally, these models require data in many, short lists for parameter estimation. However, meta-analysis of microarrays is a problem in which data are typically available only in a few, long lists. Therefore, much of the developed statistical methodology is not applicable in our setting.

To address the problem of few, long lists, we draw on algorithms developed for meta-search (Dwork et al, 2001), which is the combination of ranked results from multiple internet search engines. These algorithms conceptualize the aggregate ranking as a consensus ranking i.e., one that summarizes majority preferences between pairs of items (or genes) across lists (experiments). Different concepts of majority preference lead to different algorithms. We consider two different concepts of majority preference and evaluate the performance of the resulting algorithms in a number of simple examples and simulation studies.

The meta-search algorithm of Dwork et al represents a Markov process approach to rank aggregation and consists of two steps. First, preferences among pairs of genes are expressed in terms of a $J \times J$ transition matrix, where J is the total number of genes evaluated. The Markov process corresponding to this pairwise transition matrix has J states, where each state represents a gene. The probabilities encoded in this process's stationary distribution reflect the time spent by the process in each state. The aggregate ranking is derived by computing the process's stationary distribution; states with higher stationary probabilities are preferred to those with lower stationary probabilities and receive a higher aggregate rank.

The Markov chain (MC) approach used in the meta-search application has also been used to model global decision behavior in large groups of decision makers, such as commmuters making decisions about departure times and routing and viewers selecting from a number of television channels. In this setting (De Smet et al, 2002), multicriteria decision analysis is used to obtain a preference matrix representing the general preference structure of the group. The assumption that decision makers continously compare and re-compare pairs of alternatives during their decision process leads naturally to a Markov Chain representation. An appropriate transition matrix is derived to represent potential transitions between alternative decisions, and the limiting equilibrium distribution is interpreted as representing the global decision behavior.

In addition to the Markov process algorithms, we also develop an implementation of Thurstone's model. We show that when all studies rank the same genes, the Markov algorithms perform similarly to each other and to Thurstone's algorithm, and that all algorithms produce aggregate rankings that closely approximate the true central ranking as the number of studies increases.

An advantage of the MC algorithms is that they do not require that all lists rank the same items. This is particularly useful in the microarray setting where different studies may use microarray chips that cover overlapping, but not identical, subsets of genes.

We use our algorithms to aggregate across five published microarray studies comparing prostate tumors with normal prostate tissue (Rhodes et al, 2002). The common goal of these studies was to identify a list of genes that are most over- or under-expressed in prostate cancer and that may ultimately be useful for diagnostic or prognostic purposes or as targets for prevention or treatment. Two of these studies used Affymetrix©chips and three used custom cDNA microarrays. This dataset therefore reflects the major features that our approach is designed to address: few, long lists of genes, different technologies leading to different sets of genes within experiments, and non-comparable measurements of gene expression. We conclude with a discussion of the issue of differential variance or reliability across studies and propose methods for adjusting our approach to take study variability into account.

2 Methods

2.1 Review: Rankings, orderings, and distance measures

Assume that our goal is to aggregate across L studies or ranked lists, denoted $D_1, D_2, \ldots, D_L$. At this point we define the aggregate ranking loosely as the ranking that is closest in some sense to the individual lists. We define the notion of closeness in greater detail later in this section. The length of list D_l is denoted n_l. In what follows, we assume that each element of a list corresponds to an identifiable gene or expressed sequence tag (EST). Further, we will assume that the genes have been labelled using an annotation system that is consistent across studies. We will denote the gene labels for the genes in list l by the index $d = 1, 2, \ldots, n_l$, however, to avoid confusion in our hypothetical examples, unique items will be labelled with different letters. Let $\tau_l(d)$ denote the rank of item d in list l. Then τ_l represents a permutation of list l. Our goal is to derive an aggregate, or consensus ranking that summarizes $\tau_1, \tau_2, \ldots, \tau_L$.

In practice, we consider a reduced version of the aggregation problem, namely aggregating across top-k lists. The top-k kist is the sublist consisting of the k most highly ranked items in the original list; it is thus a *partial* list as opposed to the original, *full* list. The reason for considering top-k lists is that typically only the few most promising genes or markers can be further investigated in follow-up studies. While the number of genes in an array may be on the order of 10,000 or more, k will typically be on the order of 25 to 100. The problem of comparing and aggregating across top-k lists has been considered in the context of meta-search (Fagin et al, 2003; Dwork et al, 2001). Fagin et al defined a set of distance measures that could be used to quantify dissimilarities between top-k lists. Dwork et al considered the problem of aggregating across top-k lists. They first summarized pairwise majority preferences across top-k lists and then used the matrix of pairwise preferences to produce an MC transition matrix. The aggregate ranking was defined according to the stationary distribution of this MC. De Smet et al (2002) used a simliar approach to model the aggregate behavior of a large number of decision makers.

To establish notation for the MC approach, let U denote the union of the top-k lists. Relabel the distinct items in U from 1 to J. Let $M = \{m_{ij}\}$ be the transition matrix that reflects the preference for item i relative to item j across lists. Different preference concepts will lead to different matrices M and their corresponding stationary distributions. Assume that M is constructed to ensure a stationary distribution, $P = \{p_i\}$, $i = 1, 2, \ldots, J$ corresponding to M. The aggregate ranking is defined as the ranking that reflects the ordering of the elements of P. The element with the highest value in P receives the highest aggregate rank.

The development and evaluation of aggregate rankings requires a concept of distance between ranked lists. One well-known distance measure is Kendall's tau (Fagin et al). Kendall's tau is equal to the number of adjacent pairwise exchanges needed to convert one ranking or permutation to another. Formally, if we consider two permutations τ and τ' of a set of items U, then Kendall's tau is given by $K(\tau, \tau') = \sum_{\{i,j\} \in U} \bar{K}_{i,j}(\tau, \tau')$, where $\bar{K}_{i,j}(\tau, \tau')$ is equal to 0 if the orderings of the ranks of items i and j agree in the two lists and 1 otherwise; for example, $\tau(i) > \tau(j)$ and $\tau'(i) > \tau'(j)$ implies $\bar{K}_{i,j}(\tau, \tau') = 0$. The maximal value of Kendall's tau occurs when τ is the the reverse of τ', and this maximal value is given by $J(J-1)/2$, where J is the length of the lists. In presenting results concerning the performance of our aggregation algorithms, we use Kendall's tau and normalize all measured distances by their maximal values.

When comparing top-k lists, the aforementioned concepts of distance must be extended to handle the comparison of partial lists or, more generally, lists which rank overlapping but not identical sets of items. We use the extension of Kendall's tau suggested by Fagin et al (2003), which we briefly review here.

Consider two top-k lists, τ_1 and τ_2, which are permutations of two sets, D_1 and D_2, respectively. Let $P(\tau_1, \tau_2)$ be the set of all unordered pairs of distinct elements in $D_1 \cup D_2$. Define a penalty $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2)$ for each $i, j \in P(\tau_1, \tau_2)$. There are four possible cases:

1. **Items i and j appear in both top-k lists:** Let $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2)$ equal 0 if i and j are similarly ordered in the two lists and 1 otherwise.

2. **Items i and j appear in τ_1 and item i (but not item j) appears in τ_2:** Let $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2)$ equal zero if $\tau_1(i) < \tau_1(j)$ and 1 otherwise. Here we infer that $\tau_2(i) < \tau_2(j)$ as item i appears in τ_2 and item j does not.

3. **Item i (but not item j) appears in τ_1 and item j (but not item i) appears in τ_2** Let $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2)$ equal 1. Again, we infer the position of the missing items.

4. **Items i and j appear in τ_1 and neither i nor j appear in τ_2** Let $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2)$ equal p. In this case, we cannot infer the ordering of i and j in τ_2.

Similar to the Kendall distance, we define the top-k list Kendall distance as:

$$K^{(p)}(\tau_1, \tau_2) = \sum_{i,j \in P(\tau_1, \tau_2)} \bar{K}^{(p)}_{i,j}(\tau_1, \tau_2). \tag{1}$$

A non-zero penalty parameter, p, corresponds to the assignment of a non-zero penalty when information is missing about the ordering of i and j in one list (case 4 above). Fagin et al suggest two values for p: a neutral 0.5 or an optimistic 0. In presenting results concerning the performance of our aggregation algorithms, we set the penalty parameter to 0, which is equivalent to ignoring the relative ranking of items that are ranked lower than the kth item.

2.2 Algorithms for aggregating ranked lists

In this section we summarize three algorithms for aggregating ranked lists. All of the algorithms may be decribed as relational approaches as they are based on pairwise comparisons of items appearing in at least one of the top-k lists under study. The two MC algorithms use the results of these comparisons to construct ergodic transition matrices. The algorithm of Thurstone uses the paired comparisons to estimate an assumed continuous latent mean for each item; the aggregate ranking is then based on the order of the underlying means.

2.2.1 The MC4 algorithm

The MC4 algorithm was one of four MC algorithms developed by Dwork et al. The goal of this particular algorithm was to combat search engine spamming (manipulation of search engines to increase the chance of a web page being ranked highly). Search engines affected by spam typically have irrelevant pages ranked highly. The goal of the MC4 algorithm is to produce an aggregate ranking that ignores items (pages) that are spuriously highly ranked in only a minority of lists. The algorithm may be summarized as follows:

1. Construct the set U that consists of all items that appear within the top-k in at least one list.

2. For each pair of items i and j in U, let the preference for j over i, m^*_{ij}, equal 1 if the majority ($\geq$50%) of lists that rank both i and j rank j above i and 0 otherwise. Let $m^*_{ij} = m^*_{ji} = 0.5$ if items i and j are never directly compared in any list.

3. Define the transition matrix $M = \{m_{ij}\}$ as follows: for $i \neq j$ set m_{ij} to $m^*_{ij}/|U|$ and let $m_{ii} = 1 - \sum_{j \neq i} m_{ij}$.

4. Make the transition matrix M ergodic by multiplying each element by $1 - \varepsilon$ and then adding $\varepsilon/|U|$ to each element, where ε is a small, positive number. In practice, we use ε=0.15.

The MC4 algorithm constructs preferences based on a simple majority vote. Thus, for example, if 5 lists rank both i and j, the MC4 algorithm will produce the same value for m_{ij} regardless of whether i is preferred to j 3 out of 5 or 5 out of 5 times. This is the key to the algorithm's spam-fighting property.

2.2.2 The MCT algorithm

In contrast to the MC4, the second algorithm uses information on the frequency of the i versus j preferences. We label it MCT because this information is also used in Thurstone's order-statistics algorithm (see below). The steps in the MCT algorithm are as follows:

1. Construct the set U that consists of all items that appear within the top-k items in at least one list.

2. For each pair of items i and j in U, let the preference for j over i be $m^*_{ij} = r_{ij}/n_{ij}$, where n_{ij} is the number of lists that rank both i and j and r_{ij} is the number of these lists that rank j above i. Let $m^*_{ij} = m^*_{ji} = 0.5$ if items i and j are never directly compared in any list.

3. Define the transition matrix $M = \{m_{ij}\}$ as follows: for $i \neq j$ set m_{ij} to $m^*_{ij}/|U|$ and let $m_{ii} = 1 - \sum_{j \neq i} m_{ij}$.

4. Make the transition matrix M ergodic by multiplying each element by $1 - \varepsilon$ and then adding $\varepsilon/|U|$ to each element.

In both MC algorithms, the stationary distribution is computed by iteratively multiplying a uniform probability vector of length $|U|$ by the transition matrix. This distribution can also be computed by identifying the eigenvector associated with the eigenvalue of 1 for the transition matrix.

2.2.3 Thurstone's order-statistics algorithm

Thurstone's model for paired comparisons (Thurstone, 1927) assumes that each item has a normal distribution over a single underlying continuum and that an observer ranking any two objects is sampling from the support of this unobserved bivariate normal, ranking the object with a greater sample value above the other. In the context of gene expression levels or ratios of expression levels, the underlying continuum has a clear interpretation, and each microarray then represents a separate sampling from the multivariate normal distribution of the levels for all the genes in the chip. Consider first the measurement of just two genes, with observed values of x_1 and x_2, underlying means μ_i $(i = 1, 2)$ and variances σ_i^2 $(i = 1, 2)$, and covariance σ_{12}. The probability of $X_1 > X_2$ is then given by $\Pr(X_1 > X_2) = \Pr(X_1 - X_2 > 0)$ where $X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2 - 2\sigma_{12})$. Theoretically, f, the frequency with which gene 1 is ranked above gene 2, is calculated as:

$$f = \Phi\left(\frac{\mu_1 - \mu_2}{\left(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}\right)^{1/2}}\right) \tag{2}$$

where Φ is the normal cumulative distribution function. For situations with more than two genes, one can construct an analogous equation for each unique pair; solving this system of equations yields estimates of the parameters $\underline{\mu}$ and $\mathbf{\Sigma}$. In practice we set all variances to 1 and covariances to 0.

Our observations consist of the frequencies with which gene i is ranked above gene j for all i and j. We use a nonlinear least squares approach (Maydeu-Olivares, 1999) to estimate the parameters $\mu_2, \mu_3, \ldots, \mu_n$, substituting the observed pairwise frequencies in the relevant version of f. This substitution results in a system of $\binom{n}{2}$ equations of the form given in Eq. 2. We set the first mean, μ_1, to zero due to location indeterminacy. To avoid infinite estimates for $\underline{\mu}$, observed frequencies of 0 or 1 are adjusted by adding or subtracting a small value, which is chosen to be proportional to the number of lists that directly compare the genes. For example, if gene i and gene j are compared in n lists and i is preferred to j in all n lists, the corresponding frequencies of 0 and 1 will be adusted to $0.5/n$ and $(n-0.5)/n$, respectively. This approach ensures that the adjustment of the observed frequency away from 0 or 1 will be smaller when the observation is more reliable, i.e., based on a larger number of comparisons, and vice versa. The aggregate ranking of the genes corresponds to the rankings of the estimated parameters, $\mu_2, \mu_3, \ldots, \mu_n$.

2.3 Performance of the algorithms on two simple examples

2.3.1 A transitivity example

Consider a set of three genes $\{a, b, z\}$. Assume that the data provided to each algorithm consist of N_1 lists comparing a to z, with a preferred over z 90% of the time, and N_2 lists comparing b to z, with b preferred over z only 70% of the time.

Steps 1-3 of the MC4 algorithm yield the following transition matrix M (to two decimal places):

$$\begin{pmatrix} 0.83 & 0.17 & 0 \\ 0.17 & 0.83 & 0 \\ 0.33 & 0.33 & 0.33 \end{pmatrix},$$

with corresponding stationary distribution $(0.5, 0.5, 0)$ for items (a, b, z). In contrast, steps 1-3 of the MCT algorithm yield the transition matrix:

$$\begin{pmatrix} 0.80 & 0.17 & 0.03 \\ 0.17 & 0.73 & 0.10 \\ 0.30 & 0.23 & 0.47 \end{pmatrix},$$

with corresponding stationary distribution $(0.49, 0.4, 0.11)$ for items (a, b, z).

Since the MC4 algorithm calculates preferences based on majority, the resulting stationary distribution ranks a and b equally, as both are preferred to z by a majority of lists. In the case of the MCT algorithm, however, preferences correspond directly to the proportion of lists which prefer a over z (90%) and b over z (70%), allowing the algorithm to infer that a should be ranked more highly than b in the aggregate. Similarly, results of fitting Thurstone's model indicate that a is roughly 1.3 standard deviations above z and b is roughly 0.5 standard deviations above z. Thus, the relative ordering of a and b is determined to be that a is roughly 0.8 standard deviations above b, even though they are never directly compared, i.e., they never appear in the same list. This extended transitivity property (if a is strongly preferred to z and b is preferred to z, then a is preferred to b) is useful in combining results across microarray experiments because it enables aggregation across lists that do not contain identical genes. The MCT and Thurstone algorithms are more likely than the

MC4 algorithm to produce results that reflect this property because they use more detailed information on the observed frequencies of the various pairwise orderings in the data.

2.3.2 A signal versus noise example

To explore the ability of the algorithms to separate signal from noise, we considered 100 lists of the same 10 genes, of which some fraction of the lists were randomly ordered and the rest were consistently and correctly ordered. Over 100 such trials, we calculated the average number of correctly ranked genes produced by the algorithms as the fraction of random lists increased from 5% to 95% (see Figure 1). MC4 is the most effective algorithm for producing correctly ordered aggregates when some lists contain noise. Clearly, this is a result of using majority rule to determine pairwise preferences. The MC4 algorithm averages 10 correctly ranked genes even when 70% of lists passed to the algorithm are randomly ordered. In contrast, MCT and Thurstone's algorithm average around 6 correctly ranked genes at the same level of noise.

2.4 A simulation study

To evaluate the statistical properties of the MC algorithms in comparison with each other and with Thurstone's model, we simulated sets of microarray data from a known central ranking. We show that for simulated data, the algorithms perform similarly and produce aggregate rankings that approach the central ranking as the number of studies increases.

The data were generated as follows (Kooperberg et al, 2005): let x_{ijml} represent the expression level corresponding to the ith gene from the jth array of the mth group of the lth study for $i = 1, 2, \ldots, 100$, $j = 1, 2, \ldots, J$, $m = 1, 2$, and $l = 1, 2, \ldots, L$. Here the group indicator, m, corresponds to the two biological conditions being compared by hybridization, such as normal tissue ($m = 1$) and cancer tissue ($m = 2$). Now, generate x_{ijml} as $\mu_i + \delta_{im} + Z_{ijml}$, where $\mu_i \sim U(0, 1)$, $Z_{ijml} \sim N(0, \sigma_{il})$, $\sigma_{il} = (0.3 - 0.02\mu_i)G_{il}$, and $G_{il} \sim \Gamma(5, 1)$. Thus, the variance parameter, σ_{il}, depends on the mean, and genes with smaller expression have a larger variance; in addition, the variance parameters also vary across studies. We arbitrarily consider the first 40 genes to be of interest (truly differentially expressed). More precisely, the differential expression parameters $\delta_{im},\ m = 1, 2,\ i = 1, 2, \ldots, 100$, are set to 0 when $m = 1$ and $i = 41, 42, \ldots, 100$, and to $0.2(2B_i - 1)G_i$ otherwise, where $B_i \sim Bern(0.5)$ and $G_i \sim \Gamma(5, 1)$. The (true) central ranking corresponds to the ordering of $\delta_{1,2}, \delta_{2,2}, \ldots, \delta_{40,2}$ from largest to smallest.

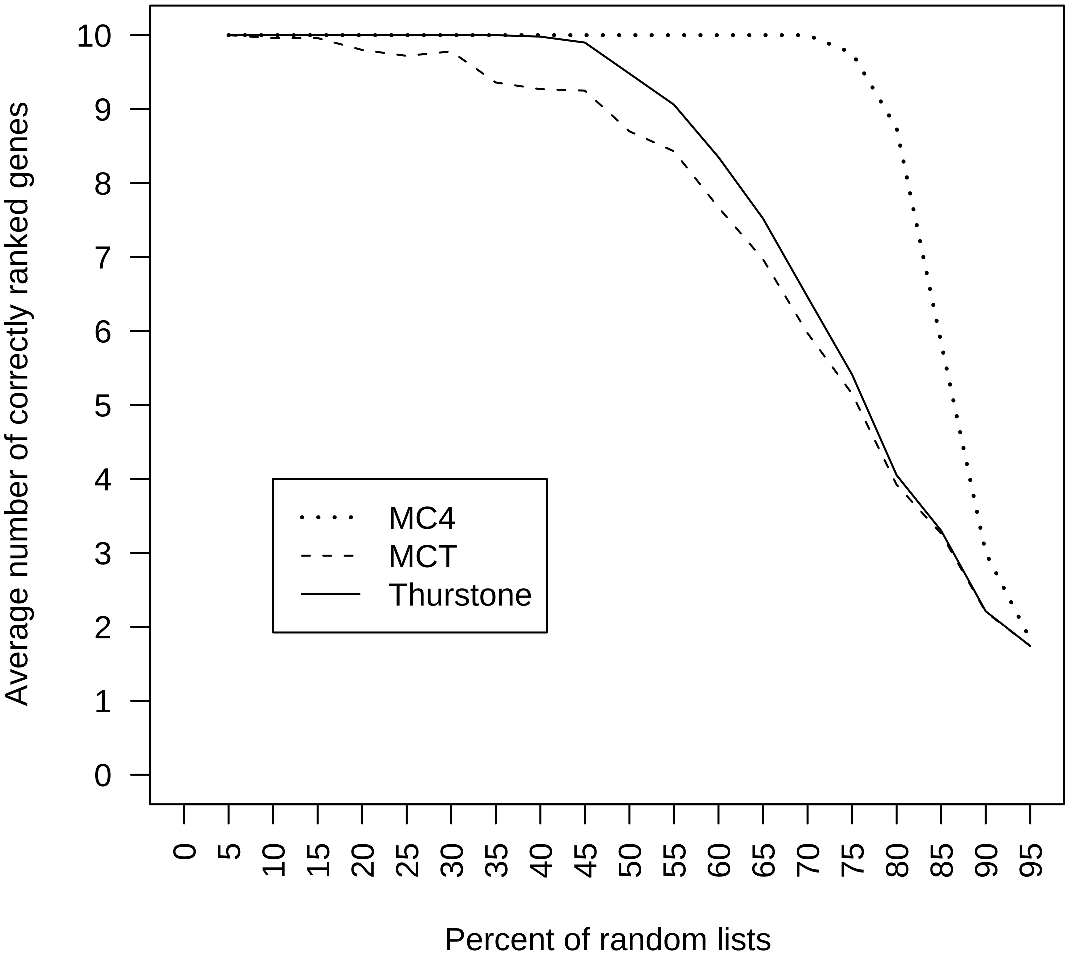

Figure 1: Over 100 trials, we passed each algorithm 100 lists of the same 10 genes, of which some fraction of the lists were randomly ordered and the rest were consistently and correctly ordered. Here we show the average number of genes ranked correctly by each algorithm as the fraction of random lists increases. MC4 is the most effective algorithm for fighting spam.

To assess the behavior of the aggregation algorithms as sample size increased, we varied L, the number of studies, over selected values from 5 to 40. We ranked the genes in each simulated study by the attenuated two-sample t-test statistic proposed by Tusher et al (2001). We also varied the number of arrays in each study, J, from 5 to 20. For each (J, L) combination, we simulated 10 datasets. Performance statistics were averaged over these 10 replicate datasets. The ranked lists from each dataset were aggregated across studies using both the MC algorithms and Thurstone's model. We set k, the number of genes of interest in the aggregate, to 40.

The top-40 genes from each aggregate were compared to the top-40 genes of the (true) central ranking in two ways. First, we computed the average distance (over the 10 simulated datasets) between the aggregate ranking and the (true) ranking using the partial-list Kendall distance measure of Fagin et al (2003). Second, we computed the number of (truly) differentially expressed genes on average in the aggregate top-40 lists produced by the various algorithms.

To facilitate comparisons across aggregation algorithms, we normalized the partial-list Kendall distance to the interval $[0, 1]$ by dividing by its maximum value, k^2. The maximum value occurs when τ_1 and τ_2 are completely disjoint. Since we set the penalty parameter p to 0, we have $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2) = 1$ only for pairs of items i, j where i appears in τ_1 and j appears in τ_2. For any two disjoint top-k lists, there are k^2 such pairs. The scale of the normalized distance can be understood by considering the following two aspects of top-k list agreement: agreement of items and agreement of rank. A normalized partial-list Kendall distance of 0 corresponds to the scenario where $\tau_1 = \tau_2$; these lists have perfect agreement of items and perfect agreement of rank. A normalized distance of 1 corresponds to the scenario where $\tau_1 \cap \tau_2 = \emptyset$; these lists have neither agreement of items nor agreement of rank. A third special case occurs when τ_1 and τ_2 include the same k items in opposite order; these lists have agreement of items and disagreement of rank. Conceptually, this scenario falls midway between perfect match lists and completely disjoint lists. Hence, we consider the normalized distance to have a reasonable scale if the third scenario results in a normalized distance equal to 0.5. The partial list Kendall distance for the third scenario is $k(k-1)/2$, as there are $k(k-1)/2$ pairs of items each with $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2) = 1$, as they appear in opposite order in τ_1 and τ_2. Thus, the normalized Kendall distance is $1/2 - 1/2k$, which asymptotically approaches 0.5, as k increases. When k equals 40, as in our simulations, the third scenario receives a normalized Kendall distance of 0.4875.

The simulation results are summarized in Tables 1 and 2. In general, the three algorithms perform similarly on the simulated data. All three algorithms exhibit the desirable property of decreasing average distance from the central ranking as the number of experiments increases. As expected, the accuracy of the aggregate solution is greater when the number of arrays per study is larger.

For our simulations, the average distance from the central ranking to the aggregates ranged from 0.03 to 0.15. The smaller distances are consistent with orderings that are generally correct with *only* a few genes incorrectly ordered. For example, let τ_1 be the correct ranking of the top-40 genes in our simulation. Transpose 16 adjacent pairs of items to produce τ_2. The normalized distance between τ_1 and τ_2 is 0.01. Larger distances are consistent with orderings that contain a few genes from outside of the true top-40. For example, replace elements $1, 11, 14, 17, 38$ and 40 of τ_1 with genes that are not contained in the top-40 to produce τ_2. The normalized distance between τ_1 and τ_2 is 0.145.

L	J	MC4	MCT	Thurstone
5	5	0.14 (0.023)	0.15 (0.017)	0.14 (0.022)
10	5	0.09 (0.008)	0.09 (0.012)	0.10 (0.011)
20	5	0.07 (0.011)	0.06 (0.009)	0.07 (0.008)
40	5	0.05 (0.008)	0.05 (0.008)	0.05 (0.009)
5	20	0.06 (0.006)	0.07 (0.009)	0.06 (0.008)
10	20	0.05 (0.008)	0.05 (0.009)	0.05 (0.008)
20	20	0.04 (0.007)	0.04 (0.005)	0.04 (0.004)
40	20	0.03 (0.005)	0.03 (0.005)	0.03 (0.006)

Table 1: Simulation results: The mean (sd) normalized Kendall distance from the known central ranking to the aggregate. The mean distance decreases as L, the number of studies, and J, the number of arrays, increase, indicating the algorithm is approaching the central ranking.

2.5 Data analysis

We used the three algorithms to aggregate results from five prostate cancer microarray studies, each of which compared normal to cancerous tissue. Three of the studies used custom-built spotted cDNA arrays (Dhanasekaran et al, 2001; Luo et al, 2001; True et al, 2006), the remaining two (Welsh et al, 2001; Singh et al, 2002) used commercially produced oligonucleotide arrays (Affymetrix©, version U95a and U95Av2). The normalized gene expression values for three of the datasets (Dhanasekaran, Luo, Welsh) were obtained by contacting Dr. Arul Chinnaiyan, a creator of the Oncomine gene expression database (www.oncomine.org). The normalized gene expression values for the remaining datasets are publicly available through the journals in which they are published. In each study, the researchers compared benign prostate tissue to prostate cancer tissue. Although metastatic prostate cancer samples were also analyzed, we limited our analysis to the comparison of clinically localized prostate cancer tissue and benign prostate tissue. Benign prostate tissue includes both normal prostate tissue and benign prostatic hyperplasia.

The five datasets consisted of different sets of array elements, with different annotation protocols. To combine the data, we standardized annotations. Genbank numbers for each array element were assigned a UniGene number using UniGene Build 180 (for details see www.ncbi.nlm.nih.gov). When multiple elements annotated to the same UniGene number, the median expression value was assigned to the element. For Affymetrix©data, array elements identified solely by identifiers from The Institute for Genomic Research (a non-profit genome research center) were removed from the analysis. Genes, defined by a unique UniGene number, which did not appear in all five studies were also removed from the analysis, leaving a total of 747 genes. Table 3 gives a brief description of the datasets.

We ranked the genes from each study by differential expression of cancer tissue relative to normal tissue, using the test statistic proposed by Tusher et al (2001). This test statistic is an attenuated t-statistic, where a small value is added to the standard deviation of each gene to help reduce the occurrence of statistically significant but biologically unimportant genes, i.e., statistically significant genes with a very low level of differential expression. The amount of overlap among the top 25 genes from each study is surprisingly small. Of the 89 genes that appear in the top-25 up-regulated genes in at least one list, only 23 appear in more than one list and only one gene, hepsin, appears in all five lists. There are three genes that appear in four lists, AMACR, GDF15, and NME1. (See Table 4).

3 Results

3.0.1 Data Analysis

Table 5 shows the results of applying our three rank aggregation algorithms to the five prostate cancer datasets. The aggregation procedures are applied to the set of genes consisting of the union, U, of the top-25 genes from each study (89 up-regulated genes). Unless otherwise stated, results presented are based on 100 iterations of the MC. Table 4 gives the observed top-25 upregulated genes for each of the five studies. Table 5 gives the top-25 up-regulated genes as determined by each of the three algorithms. The table also gives the stationary distribution for the two MC algorithms and the estimated μ values for the Thurstone algorithm. The three algorithms give quite similar results: of the aggregate top-25 up-regulated genes, 25 are

L	J	MC4	MCT	Thurstone
5	5	32	32	32
10	5	34	34	34
20	5	36	35	36
40	5	36	36	36
5	20	37	38	37
10	20	39	39	38
20	20	39	39	39
40	20	40	40	40

Table 2: Simulation results: The number of true discoveries, i.e. genes which appear in the top-40 in the central (true) ranking which also appear in the top-40 of the aggregate. The number of true discoveries increases as L, the number of studies, and J, the number of arrays, increases, indicating the algorithm is approaching the central ranking.

Authors	Number of Clones	Number of Samples	
		BP	CaP
Dhanasekaran et al	7150	22	59
Luo et al	5831	9	16
Welsh et al	7567	9	25
True et al	4653	32*	32*
Singh et al	6621	50	52

Table 3: Description of Datasets used for meta-analysis. BP:Benign Prostate, CaP:Localized Prostate Cancer. * The experiment by True et al hybridized matched localized cancer tissue and benign prostate tissue in a head-to-head fashion. There were a total of 32 cases in this experiment.

identified by all three algorithms. Of the two MC algorithms, the MCT algorithm tends to be more consistent with Thurstone's method.

While the dimensionality of the problem precludes us from fully detailing how the aggregate rankings are produced, some intuition may be gained from studying the genes that are most highly ranked in the aggregates. Among the up-regulated genes, the most highly ranked are HPN, AMACR, NME1 and GDF15. HPN is the only gene that appears in the top-25 lists from all five studies, at positions 1, 1, 4, 2 and 1 in the studies of Luo, Welsh, Dhanasekharan, True, and Singh respectively. It is followed in the aggregate ranking(s) by AMACR, which appears at positions 2, 2, 2, 1 and 38. It is instructive to study pairs of genes that are ordered differently by the MC4 algorithm than by the other algorithms. For example, the MC4 algorithm obtains quite similar stationary probabilities for GDF15 and NME1, with GDF15 just above NME1, but the MCT and Thurstone algorithms consistently rank NME1 above GDF15. To explain this, note that GDF15 ranks higher than NME1 in four out of five studies, and the "majority rule" MC4 algorithm yields a result consistent with this observation. However, in the fifth study (Luo et al), GDF15 appears at position 66, far below the number 15 slot of NME1. Since the MCT and Thurstone algorithms will reduce the preference for GDF15 in relation to all the genes appearing between NME1 and GDF15 (ranks 16-65), this poor performance by GDF15 in a single study is sufficient to bring down the final position of this gene in the aggregate. A simliar phenomenon is observed when we consider the genes SND1 and FASN; FASN is ranked more highly than SND1 by the MC4 algorithm, but the reverse is true for the other two algorithms, which rank FASN considerably lower than SND1. Again we note that even though FASN outranks SND1 in three out of five lists, FASN appears at positions 82 and 79 in the Luo and True studies, whereas the lowest rank received by SND1 is 36 in the study of Luo et al. Thus, even if a gene is fairly highly ranked in three or more studies, a low ranking in one or two studies is enough to move it lower in the aggregate, and this is more likely to occur when aggregating via the MCT and Thurstone algorithms than via the MC4 algorithm. Also, note that although OGT is the most highly-ranked up-regulated gene in the study of Dhanasekharan et al, it does not appear in any of the aggregates in Table 4. This is because OGT appears at positions 83, 56 and 55 in the studies of Luo, Welsh and Singh and at position 40 in the study of True.

Among the highly-ranked aggregrate results are several genes that have already been identified as important in prostate cancer development and progression, including hepsin (HPN) which stimulates metastasis formation in an animal model of prostate cancer (Klezovitch et al, 2004), alpha-Methylacyl-CoA racemase (AMACR), a clinically utilized marker of prostate cancer (Kuefer et al, 2002), and fatty acid synthase (FASN), an emerging therapeutic target (Pizer et al, 2001). The up-regulated gene list also included several genes linked to signal transduction and gene transcription including GUCY1A3, an androgen-receptor-regulated guanylate cyclase implicated in prostate carcinogeneses (Dong et al, 2005), ANK3, a member of the ankyrin family of structural proteins (Ignatiuk et al, 2006) and STRA13, a basic helix-loop-helix (bHLH) transcription factor which regulates cell differentiation, proliferation, apoptosis and the response to hypoxic conditions (Ivanova et al, 2005). The up-regulated results also include chaperone genes. The proteins encoded by these genes facilitate the folding of newly synthesized proteins in the endoplasmic reticulum (CCT2, CANX or calnexin) or stabilize proteins in the cytoplasm (TRAP1).

4 Discussion

In this article we have proposed a method for aggregating results across microarray experiments. Our approach draws on methods from the meta-search and multicriteria decision-making literatures and relies on information on the rankings of genes within each experiment rather than on quantitative measures of gene expression. The utility of the rank metric was highlighted recently by Xu et al (2005) who developed a rank-based approach for classifying tissue samples and applied it to gene-expression profiles from multiple experiments. In fact, the classifier of Xu et al was based on the relative rankings of pairs of genes within the classes of interest. This work and ours confirm the findings of Yuen et al (2002), namely that rank-based information can yield robust inferences across microarray studies.

Our proposed approach consists of two distinct methods for rank aggregation, the first being algorithmic and based on the methods of Dwork et al (2001) in the computer science literature and the second being statistical and based on a long-standing estimating algorithm from the statistics literature. We have built on and extended the Dwork methodology by introducing a variant on the MC4 algorithm (the MCT algorithm) and examinining the properties of both these algorithms in several illustrative examples and a simulation study. The examples highlight the differences between the algorithms, particularly the spam-fighting property of the MC4 algorithm and the extended transitivity property of the other two. Our results indicate that even though the MC algorithms are heuristic, they still display desirable statistical properties as the sample size increases. In our simulations and sample dataset, the MC algorithms produce results similar to Thurstone's; this is another validation of the performance of the MC approach since the Thurstone approach is the most rigorous statistically.

The rank aggregation approach has both advantages and limitations relative to other meta-analysis methods in the setting of gene-expression studies. In contrast to other methods, our approach does not require a pre-processing step to identify a set of genes common to all studies. In our example, we did limit attention to this common set of genes to avoid the setting where very few of the top-25 results overlapped across studies. We conducted a supplementary analysis, which included all genes common to at least three out of the five studies, and were able to obtain aggregate rankings for this expanded set of genes. While the top-25 upregulated lists still had HPN and AMACR in the top two positions, these were followed by several new genes that preceded NLM1 and GDF15 in the aggregate orderings, namely, HSPD1, TARP, CAMKK2, TXN, and MY06. There was also greater heterogeneity between the aggregates computed by the different algorithms when considering the expanded set of genes. In practice, we would recommend using as large a gene superset as possible, but this must be balanced against the realization that meaningful aggregation can only occur when there is considerable overlap between the gene lists being aggregated.

Another advantage of the proposed approach is that it produces a ranked list of genes rather than a set of genes that are not distinguished by priority or preference. The availability of a ranked list can be important when selecting genes for further investigation under resource constraints. Limitations include the fact that with few lists the granularity of the preference matrix in terms of the range of its elements is quite limited. Moreover, these preference probabilities do not reflect the number of lists being considered. However, our simulations show that as the number of lists increases, inferences do tend to become more accurate.

Our approach does not explicitly consider the fact that different studies may have different levels of reliability. The standard statistical measure of reliability is precision or variance. In principle, if variance measures were available for each study, a weighted preference matrix could be obtained, where preference information from more reliable studies would be upweighted by a factor proportional to the inverse of the study-specific variance. In practice, however, a concept of variance must be developed that is appropriate for ranked lists, and this variance must then be estimated for each study. Marden (1995) defines a concept of spread for a set of ranked lists, given by the average distance between each of the lists and a central ranking. Future work will concentrate on the implementation of this concept in the setting of multiple microarray experiments and the adaptation of our aggregation procedures to take study reliability into account.

Rank	Luo	Welsh	Dhana	True	Singh
1	HPN	HPN	OGT	AMACR	HPN
2	AMACR	AMACR	AMACR	HPN	SLC25A6
3	CYP1B1	OACT2	FASN	NME2	EEF2
4	ATF5	GDF15	HPN	CBX3	SAT
5	BRCA1	FASN	UAP1	GDF15	NME2
6	LGALS3	ANK3	GUCY1A3	MTHFD2	LDHA
7	MYC	KRT18	OACT2	MRPL3	CANX
8	PCDHGC3	UAP1	SLC19A1	SLC25A6	NACA
9	WT1	GRP58	KRT18	NME1	FASN
10	TFF3	PPIB	EEF2	COX6C	SND1
11	MARCKS	KRT7	STRA13	JTV1	KRT18
12	OS−9	NME1	ALCAM	CCNG2	RPL15
13	CCND2	STRA13	GDF15	AP3S1	TNFSF10
14	NME1	DAPK1	NME1	EEF2	SERP1
15	DYRK1A	TMEM4	CALR	RAN	GRP58
16	TRAP1	CANX	SND1	PRKACA	ALCAM
17	FMO5	TRA1	STAT6	RAD23B	GDF15
18	ZHX2	PRSS8	TCEB3	PSAP	TMEM4
19	RPL36AL	ENTPD6	EIF4A1	CCT2	CCT2
20	ITPR3	PPP1CA	LMAN1	G3BP	SLC39A6
21	GCSH	ACADSB	MAOA	EPRS	RPL5
22	DDB2	PTPLB	ATP6V0B	CKAP1	RPS13
23	TFCP2	TMEM23	PPIB	LIG3	MTHFD2
24	TRAM1	MRPL3	FMO5	SNX4	G3BP2
25	YTHDF3	SLC19A1	SLC7A5	NSMAF	UAP1

Table 4: Up-regulated top-25: The observed rankings from the five studies are given.

Rank	MC4		MCT		Thurstone	
1	HPN	0.070	HPN	0.066	HPN	2.362
2	AMACR	0.062	AMACR	0.042	AMACR	1.880
3	GDF15	0.041	NME1	0.029	NME1	1.360
4	NME1	0.040	GDF15	0.024	GDF15	1.216
5	SLC25A6	0.038	EEF2	0.023	EEF2	1.150
6	KRT18	0.037	SND1	0.021	SND1	1.094
7	EEF2	0.037	KRT18	0.021	KRT18	1.052
8	FASN	0.032	SLC25A6	0.020	UAP1	1.023
9	GUCY1A3	0.031	UAP1	0.020	SLC25A6	1.019
10	SND1	0.029	GUCY1A3	0.019	GUCY1A3	0.967
11	ANK3	0.029	STRA13	0.018	STRA13	0.883
12	OACT2	0.027	OACT2	0.018	OACT2	0.868
13	UAP1	0.026	ANK3	0.017	ANK3	0.819
14	STRA13	0.021	MRPL3	0.015	MRPL3	0.654
15	MRPL3	0.019	MTHFD2	0.015	MTHFD2	0.611
16	SERP1	0.017	FASN	0.014	CCT2	0.579
17	PPIB	0.016	CCT2	0.014	PPIB	0.558
18	MTHFD2	0.015	PPIB	0.014	FASN	0.527
19	CCT2	0.015	ALCAM	0.013	ALCAM	0.512
20	ALCAM	0.014	TRAP1	0.013	TRAP1	0.507
21	MAOA	0.014	SERP1	0.013	MAOA	0.489
22	TRAP1	0.014	MAOA	0.013	SERP1	0.448
23	TFF3	0.013	TFF3	0.013	CANX	0.436
24	GRP58	0.013	CANX	0.013	TFF3	0.432
25	CANX	0.013	GRP58	0.012	GRP58	0.425

Table 5: Up-regulated top-25 aggregation results: The aggregate rankings and associated stationary distributions produced by MC4 and MCT are given in the first four columns. The stationary probabilites are rounded to 3 decimal places, resulting in the appearance of ties between some genes. There are no actual ties in the rankings. The final columns give the aggregate ranking produced by Thurstones method together with the estimated μ values for each gene.

References

Bradley, R.A. and Terry, M.A. (1952) Rank analysis of incomplete block designs. I. Biometrika, 39:324-345.

Benjamini Y., Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B, 57(1):289-300.

De Smet, Y., Sprinagael, J., Kunsch, P. (2002) Towards statistical multicritera decision modeling: A first approach. Journal of Multi-Criteria Decision Analyis, 11(6):305-313.

Dhanasekaran, S.M., Barrette, T.R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K.J., Rubin, M.S. and Chinnaiyan,A.M. (2001) Delineation of prognostic biomarkers in prostate cancer. Nature, 412(6839):822-826.

Dong, Y., Zhang, H., Gao, A.C., Marshall, J.R., Ip, C. (2005) Androgen receptor signaling intensity is a key factor in determining the sensitivity of prostate cancer cells to selenium inhibition of growth and cancer-specific biomarkers. Mol Cancer Ther 4(7):1047-1055.

Dwork, C., Kumar, R., Naor, M. and Sivakumar, D. (2001) Rank aggregation methods for the web. http://www10.org/cdrom/papers/577/

Fagin, R., Kumar, R. and Sivakumar, D. (2003) Comparing top-k lists. SIAM J. Discrete Math., 17(1):134-160.

Fligner, M.A. and Verducci,J.S. (1988) Multistage ranking models. J Am Stat Assoc., 83(403):892-901.

Ignatiuk, A., Quickfall, J.P., Hawrysh, A.D., Camberlain, M.D., Anderson, D.H. (2006) The smaller isoforms of ankyrin 3 bind to the p85 subunit of phosphatidylinositol 3’-kinase and enhance platelet-derived growth factor receptor down-regulation. J Biol Chem 281(9):5956-5964.

Ivanova, A., Liao, S.Y., Lerman, M.I., Ivanov, S., Stanbridge, F.J. (2005) STRA13 expression and subcellular localisation in normal and tumour tissues: implications for use as a diagnostic and differentiaion marker. J Med Genet 42(7):556-576.

Klezovitch, O., Chevillet, J., Mirosevich, J., Roberts, R., Matusik, R., Vasioukhin, V. (2004) Hepsin promotes prostate cancer and metastasis. Cancer Cell. 6(2):185-195

Kooperberg, C., Aragaki, A., Strand, A.D., Olson, J.M. (2005) Significance testing for small microarray experiments. Statistics in Medicine 24(15):2281-2298.

Kuefer, R., Varambally, S., Zhou, M., Lucas, P.C., Loeffler, M., Wolter, H., Mattfeldt, T., Hautmann, R.E., Gschwend, J.E., Barrette, T.R., Dunn, R.L., Chinnaiyan, A.M., Rubin, M.A. (2002) alpha-Methylacyl-CoA racemase: expression levels of this novel cancer biomarker depend on tumor differentiation. Am J Pathol 161(3):841-848.

Luo, J., Duggan, D.J., Chen, Y., Sauvageot, J., Ewing, .M., Bittner, M.L., Trent, J.M. and Isaacs, W.B. (2001) Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. Cancer Res., 61(12):4683-4688.

Marden, J.I. (1995) Analyzing and modeling rank data. Chapman and Hall, London.

Maydeu-Olivares, A. (1999) Thursonian modeling of ranking data via mean and covariance structure analysis. Psychometrika, 64(3):325-340.

McCarroll, S.A., Murphy, C.T., Zou, S., Pletcher, S.D., Chin, C.S., Kenyon, C., Bargmann, C. and Li, H. (2004) Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. Nature Genetics, 36(2):197-204.

Parmigiani, G., Garrett-Mayer, E.S., Anbazhagan, R. and Garielson, E. (2004) A cross-study comparison of gene expression studies for the molecular classification of lung cancer. Clinical Cancer Research, 10(9):2922-2927.

Pizer, E.S., Pflug, B.R., Bova, G.S., Han, W.F., Udan, M.S., Nelson, J.B. (2001) Increased fatty acid synthase as a therapeutic target in adrogen-independent prostate cancer progression. Prostate 47(2):102-110.

Plackett, R.L. (1975) The analysis of permutations. Applied Statistics, 24:193-202.

Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh,D. and Chinnaiyan,A.M. (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res., 62(15):4427-4433.

Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A. and Chinnaiyan, A.M. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci USA., 101(25):9309-9314.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., Sellers, W.R. (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer Cell, 1(2):203-209.

Smith, B.B. (1950) Discussion of professor Ross's paper. J R Stat Soc Ser B, 12:53-56.

Thurstone, L.L. (1927) A law of comparative judgement. Phycological Rev., 79:281-299.

Thurstone, L.L. (1931) Rank order as a psychological method. J Exp. Psychol., 14:187-201.

True, L., Coleman, I., Hawley, S., Huang, A., Gifford, D., Coleman, R., Beer, T., Gelman, E., Datta, M., Mostaghel, E., Knudsen, B., Lange, P., Vessella, R., Lin, D., Hood, L., Nelson, P. (2006) A Molecular Correlate to the Gleason Grading System for Prostate Adenocarcinoma. Proc Natl Acad Sci U S A., forthcoming.

Tusher, V.G., Tibshirani, R., Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A., 98(9):5116-21.

Varambally S., Yu J., Laxman B., Rhodes D.R., Mehra R., Tomlins S.A., Shah R.B., Chandran U., Monzon F.A., Becich M.J., Wei J.T., Pienta K.J., Ghosh D., Rubin M.A., Chinnaiyan A.M. (2005) Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. Cancer Cell. 8(5):393-406.

Wang, J., Coombes, K.R., Highsmith, W.E., Keating, M.J and Abruzzo, L.V. (2004) Differences in gene expression between B-cell chronic lymphcytic leukemia and normal B cells: a meta-analysis of three microarray studies. Bioinformatics, 20(17):3166-3178.

Welsh, J.B., Sapinoso, L.M., Su, A.I., Kern, S.G., Wang-Rodriguez, J., Moskaluk, C.A., Frierson, H.F.Jr. and Hampton,G.M. (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. Cancer Res., 61(16):5974-5978.

Xu L., Tan A.C., Naiman D.Q., Geman D., and Winslow R.L. (2005) Robust cancer marker genes emerge from direct integration of inter-study microarray data. Bioinformatics, 21(20):3905-3911.

Yuen, T., Wurmbach, E., Pfeffer, R.L., Ebersole, B.J. and Sealfon, S.C. (2002) Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. Nucleic Acids Res., 30:e48.