CBCS
http://agrobiol.sggw.waw.pl/cbcs

Communications in Biometry and Crop Science
Vol. 5, No. 1, 2010, pp. 41–55

International Journal of the Faculty of Agriculture and Biology,
Warsaw University of Life Sciences, Poland

REGULAR ARTICLE

Combining partially ranked data in plant breeding and biology: I. Rank aggregating methods

Ivan Simko*, Dov A. Pechenick

USDA-ARS, Crop Improvement and Protection Research Unit, 1636 East Alisal Street, Salinas, CA 93905, USA.
* Corresponding author: Ivan Simko, E-mail: Ivan.Simko@ars.usda.gov

CITATION: Simko, I., Pechenick, D.A (2010). Combining partially ranked data in plant breeding and biology: I. Rank aggregating methods. *Communications in Biometry and Crop Science* 5 (1), 41–55.

Received: 13 November 2009, Accepted: 20 June 2010, Published online: 15 July 2010

ABSTRACT

Combining heterogeneous data from plant breeding trials into a single dataset can be challenging, especially if observations have been performed only on partially overlapping sets of accessions, or if evaluations were done with different rating scales. In the present work we propose combining such data by making use of aggregate ranking approaches. To test 13 aggregate ranking methods for performance, we have simulated 16 types of datasets that resemble those observed in plant breeding trials. The evaluation of aggregate ranking methods was carried out using both distance-based measures (Kendall's *tau* and Spearman's *rho*) and number of rank violations caused by a proposed aggregate ranking. Our analysis indicates that methods based on Bradley-Terry or Rasch models performed better than the other tested methods when factors such as fitness of aggregate rankings, time required for analyses, and ability to analyze weak rankings were considered. Verification of the approach on real data from 19 studies indicated a substantial increase in significance (*P*-value dropped by a factor of 100,000) when linkage between a marker and a trait was based on aggregated data rather than on each of the individual trials. The ability to combine heterogeneous data from independent studies has important ramifications for data analysis in association studies. Results from our study indicate that this kind of meta-analysis is more powerful than individual analyses.

Key Words: *Adjusted means; Bradley-Terry model; Markov chains; partially ranked data; Plackett-Luce model; rank-aggregation; Rasch model; sampling methods.*

INTRODUCTION

Plant breeders test annually a large number of accessions in multiple trials, continuously generating considerable amounts of data. However, combining data from trials performed at

different locations, years, or laboratories (breeding stations) is complicated because measurements used in the evaluations are not identical. For example, potato breeders regularly evaluate 'earliness', which indicates how early tubers set on tested material. Since direct observation of this trait is usually not possible, breeders use indirect assessment of earliness, such as number or weight of tubers or percent of plants that formed tubers at a certain date. Another approach is to count the number of days until the first tuber appears on a plant, or use an arbitrary scale to indicate plant vine maturity as it relates to early tuber formation. Yet another approach uses a test of tuber formation on stem cuttings grown in a greenhouse or in vitro (Simko et al., 1999; van den Berg et al., 1996). When merging such diverse sets of measurements is desired, threshold models proposed by Hartung and Piepho (2005) can be applied. However, if distribution-free methods are preferred (e.g. because of their smaller sensitivity to errors of measurement), the absolute values might be replaced with relative ranks, and the ranks combined into a single aggregated ranking. Yet another factor that has to be considered when combining data from plant-breeding trials is that different numbers of accessions are tested in separate trials and usually only a few (if any) of them are represented in any two trials. This type of ranked data is equivalent to partially ranked data from multiple ranked lists (Cook et al., 2007).

Aggregation of partially ranked data can be done through a number of techniques that range from the simple that are based on averages to the complex that employ advanced computational methodologies (Marden, 1995). Simple rank aggregation techniques are easy to calculate, but they may not provide an optimal ranking when different subsets of accessions are analyzed in separate trials. The more complex techniques that take into consideration not only ranks, but also which accessions were compared in each trial, include the order-statistic models of Thurstone (1927; 1931), paired comparison models (Bradley and Terry, 1952), multistage models (Luce, 1959; Plackett, 1975), and methods based on Markov chains (MC) (Norris, 1997).

The objective of the present work was to test and compare suitability of different methods for aggregating partially ranked data typically found in plant-breeding trials. Evaluation of 13 rank-aggregating methods was performed on 80 computer-generated datasets and real data obtained from testing earliness of potato tuber formation.

MATERIALS AND METHODS

SIMPLE POSITIONAL METHODS

The primary advantage of simple methods is that they are easy to compute and a rank aggregation can be obtained quickly for even large datasets. These methods usually assign a score corresponding to the positions in which accessions appear in partial ranked lists, and then these scores are combined into a total score for each accession that is used to construct a final ranking. The two tested simple positional methods were RankProd (RP) (Breitling et al., 2004) and METRADISC (MD) (Zintzaras and Ioannidis, 2008).

METHODS BASED ON ADJUSTED MEANS

The fitting of models here is based on the established principle of least squares. Least squares mean is defined as a linear combination of the estimated means from a linear model (Piepho, 2003). In other words, it is the observed mean appropriately adjusted for the other effects in the model. When no missing values are present in the dataset, mean and least squares mean of the data are identical. The two tested methods were based on additive model (AD) and regression model (RG) (Piepho, 2003). Though these methods were developed for analysis of metric data, they are often used to analyze ordinal data that do not meet the usual assumptions of homogeneity of variance, normality, and linearity/additivity (Hartung and Piepho, 2005).

METHODS BASED ON PAIRED COMPARISONS

An alternative to ranking all accessions in a trial is to choose which accession from each pair of accessions is preferred (ranked higher). Once all pairwise preferences are established from a given set of rankings, methods that consider head-to-head comparisons may be employed. The Bradley-Terry model (Bradley and Terry, 1952) for paired comparisons is frequently used to calculate probabilities of the possible outcomes when accessions are judged in pairs. The model for a pair of accessions is:

$$P(\text{accession } i \text{ outperforms accession } j) = \frac{\gamma_i}{\gamma_i + \gamma_j} \quad (1)$$

where γ_i is a positive-valued parameter associated with accession i, for all of the comparisons pitting accession i against accession j. The four tested methods based on paired comparisons were ELOstat (ES) (Schubert, 2000a; 2000b), BayesELO (BE) (Coulom, 2008), Colley's algorithm (CL) (Colley, 2002), and Mease's algorithm (MS) (Mease, 2003).

METHODS BASED ON MULTISTAGE MODELS

Luce (1959) extended the Bradley-Terry model by allowing for more general comparisons than just those that are paired, and Plackett (1975) presented a saturated model from probabilities of winning:

$$P[W = w; v] = \prod_{i=1}^{m} \frac{v_{w_i}}{v_{w_i} + v_{w_{i+1}} + \cdots + v_{w_m}} \quad (2)$$

This generalization of the Bradley-Terry model was termed the Plackett-Luce model (Marden, 1995). The three tested methods based on multistage models were Hunter's algorithm (HN) (Hunter, 2004), Grave's algorithm (GR) (Graves et al., 2003), and Rasch model (RS) for multiple comparisons (Linacre, 1992).

METHODS BASED ON MARKOV CHAINS

In probability theory studies, Markov chains are used as models for random phenomena evolving in time (Norris, 1997). A Markov chain is a sequence of variables generated by a stochastic process whose future states retain no memory of past states, and depend on the present state only:

$$\Pr(X_{n+1} = x \mid X_n = x_n) \quad (3)$$

Models based on Markov chains can be used to convert pairwise preferences into a stationary distribution that can be solved for an aggregate ranking (DeConde et al., 2006). This type of model is frequently used in situations where very large numbers of accessions need to be ranked. The two tested methods based on Markov chains were PageRank (PR) (Page et al., 1998) and a modified version of the Markov chain – Thurstone model (MT) (DeConde et al., 2006). Our modification of the MT model assumes no information about the order of direct comparison between individuals that were, in fact, never compared.

SIMULATIONS

SIMULATED DATA

To evaluate and compare the statistical properties of the various ranking algorithms, we constructed sets of simulated data that resemble those observed in plant breeding trials. Two 100 × 100 full matrices were generated – each of them an equivalent of 100 accessions tested in 100 trials. For each accession in each trial the 'observed value' was generated by adding the 'mean value' for that accession and a random 'noise' component, as described in Cook et al. (2007). Mean values for all accessions were drawn from a normal distribution $N(100,225)$, while noise for the two matrices was drawn from a normal distribution $N(0,100)$ and $N(0,625)$, respectively. Because of this difference in noise, the mean of Pearson's correlation coefficient between trials was 0.67 for the first matrix and 0.25 for the second one.

SAMPLING METHODS

Four different sampling methods were employed to randomly construct incomplete datasets from the original 100-accession, 100-trial full matrices, where only a fraction of the total accessions were included in each trial (Figure 1). The rationale for the different sampling methods was to create data whose pattern would resemble those seen in plant breeding trials. For example, random sampling 'R' imitates data obtained from different breeding programs, locations, and years with minimal and inconsistent overlap, while the drop-and-replace pattern of 'C' sampling resembles performance trials where partially overlapping subsets of accessions are tested in consecutive years. Non-random samplings 'B' and 'A' emulate trials at two or more locations with incompletely overlapping subsets of accessions that differ substantially in performance. While all four methods generated these datasets in a stochastic fashion, constraints were placed on some of the sampling methods to ensure that certain rules be followed. All datasets chosen for subsequent analysis shared certain properties:

1. Each accession was tested in at least 1 trial and not allowed to be tested in all trials (with the exception of four controls in "non-random sampling C" that were tested in all 100 trials).
2. Each trial contained at least 2 and at most 99 accessions.
3. No accession performed consistently best or worst in every trial in which it was tested.
4. No two accessions had identical observed values in any one trial.

Random sampling (R): Accessions were randomly selected according to a binomial distribution, where the probability of any accession being selected in any trial was equal.

Non-random sampling (A): Accessions were randomly selected such that in no trial the expected best and worst accessions would be directly compared.

Non-random sampling (B): Accessions were randomly selected such that half the trials would contain frequent direct comparisons between the best-performing and worst-performing, and half would contain comparisons between middle-performing and worst-performing accessions, while none would contain direct comparisons between best-performing and middle-performing accessions.

Non-random sampling (C): Every trial contained four controls: the overall best two and worst two performing accessions. Beside controls, the accessions for the first trial were selected at random for a desired density of data. The same accessions were selected for inclusion in the following trial, with the exception of one accession that was dropped and replaced by a new accession. The dropped accession had to have already been included in at least as many or more trials than any of the remaining accessions. The replacement accession was selected randomly from a pool of accessions that had not yet been included in any trial, or had been included in fewer trials than any other accession. The drop-and-replace pattern of this sampling resembles an evaluation method proposed by Halekoh and Kristensen (2008).

DENSITY OF DATA-POINTS

Each sampling method was designed to reduce the number of data-points from 10,000 (100 accessions tested in 100 trials) in the full matrices to either ~20% or ~7%, hereon referred to as data-point density. Datasets with similar data-point densities had different pairwise densities (proportion of unique direct pairwise comparisons), depending on the sampling method used. Five datasets were generated for every combination of the four sampling methods, two data-point densities, and two original matrices. After generating these 80 datasets, the ranking for each trial in each dataset was determined by sorting the observed values in descending order and replacing each value by its relative position. These rankings were then used as input for the aggregation methods.

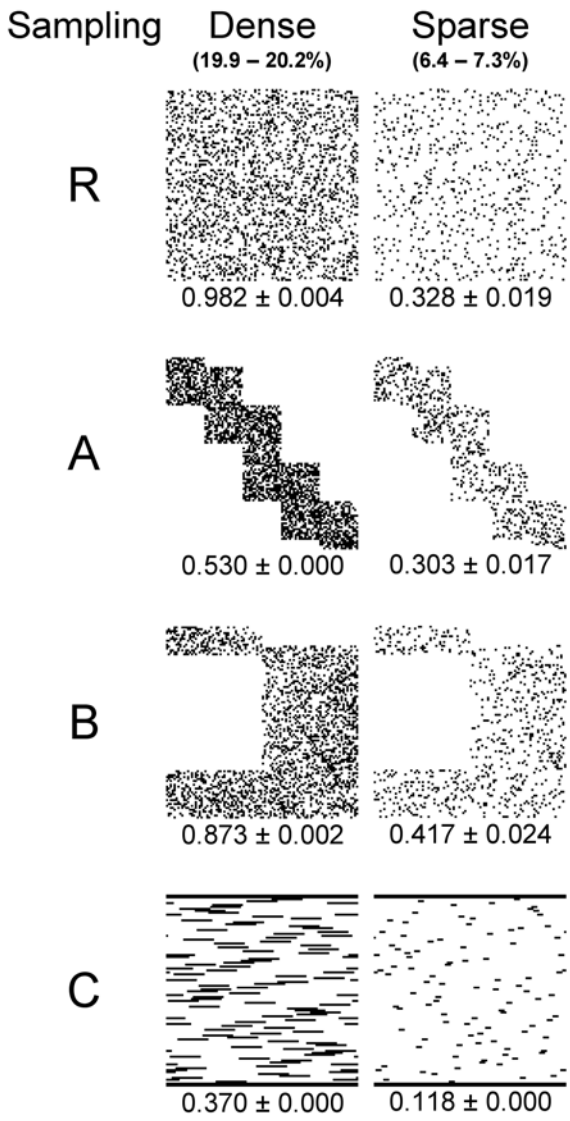


Figure 1. Demonstration of differences among four sampling methods and two densities of data-points. Each black dot shows a data-point used in aggregate ranking analyses. These data-points were sampled from an original full matrix of 100 accessions (rows) tested in 100 trials (columns). For easier illustration, accessions were sorted in descending order according to their mean performances in the original full matrix. Below each plot are listed pairwise densities.

INVERTED RANKING

Some of the tested methods (RP, HN, GR, PR, MT) are based on models whose objective is to identify only a few best performing accessions, while resolution for bottom ranked accessions is low. To increase resolution on both ends of distribution, calculations were carried out for both regular and inverted ranked lists and combined into a single aggregated ranking.

MEASURES OF PERFORMANCE

DISTANCE-BASED MEASURES AND RANK VIOLATIONS

Evaluation of performance requires some measure of the distance between the aggregate rankings determined by tested methods and an expected ranking. In real datasets the expected ranking is usually not known; however, our simulated datasets were generated from known normal distributions and therefore the means of observed values for each dataset were used to determine the expected ranking. These expected rankings were subsequently used to measure similarity to aggregate rankings with Spearman's *rho* and Kendall's *tau* rank correlation coefficients (Marden, 1995). In addition to the rank correlations, performance of all methods was evaluated by the number of rank violations caused by a proposed aggregate ranking. The number of rank violations was calculated by comparing an aggregate ranking and the ranked list of accessions in each trial from which the aggregate ranking was generated (Cook et al., 2007). A ranking violation between two accessions occurs when they are ranked differently in a trial as compared to the aggregate ranking. If the two rankings agree, no violation is accrued. The calculated number of rank violations was subsequently transformed into a proportion, by dividing it with the maximum possible number of rank violations for a given dataset.

ANALYSIS OF PERFORMANCE

In each type of dataset, one-way analysis of variance (ANOVA) was performed to test equality of the average correlation coefficients (or proportion of rank violations) among all 13 methods. If it was determined by ANOVA that means were not equal, Hsu's MCB test (Hsu, 1981) was applied to detect statistical significance between the best performing method (BPM) for the particular type of dataset and all other tested methods. Three-way ANOVA was carried out to identify factors affecting performance of a method in different types of datasets. To compare overall performance of tested methods, the receiver-operating characteristic (ROC) curve analysis (Hanley and McNeil, 1982) was carried out on 8,000 ranks from each method and the area under the ROC curve (AUC) was calculated.

ANALYSIS OF REAL DATA

To demonstrate the value of aggregate ranking on real data, we used observations of tuber formation from 157 accessions and 19 independent trials (Supplementary Table S1: http://agrobiol.sggw.waw.pl/~cbcs/articles/5_1_7/Supplementary_Table_S1.xls).

Earliness of tuber formation was indirectly assessed as number of tubers, weight of tubers, percent of tuber-forming plants, number of tubers in vitro, weight of tubers in vitro, percent of greenhouse-grown cutting with tubers, and number of days from planting to initiation of the first tuber. Detailed information about traits and scoring is in Simko et al. (1999) and van den Berg et al. (1996). A subset of data was randomly selected (through random sampling described above) from the dataset in such a way that each trial contained only 12 to 41 accessions and no accession was tested in more than eight trials. Eleven of the 13 methods were used to combine data from individual trials into aggregated rankings (HN and GR were not tested, because both methods can aggregate only strong rankings without ties).

We hypothesized that aggregate rankings are more likely to detect the true marker-trait association than limited data from individual trials. To test this hypothesis we used the molecular marker TG441 (on chromosome 5) that is strongly associated with earliness of tuberization (Simko et al., 1999, van den Berg et al., 1996) and was previously used to genotype all 157 accessions. Association between TG441 and earliness of tuberization was calculated in each of the 19 individual trials, and also for aggregated rankings produced by the 11 methods. The level of association was expressed as $-\log(P)$, where P is the probability of obtaining a test statistics as large (or larger) than observed. To examine the effect of sample size on $-\log(P)$ values, 1,000 permutations were generated from ranked data

observed in each trial and also from aggregated ranks calculated for 157 accessions. The –log(*P*) values calculated from permutated ranks were then compared to those obtained from actual aggregated rankings.

SOFTWARE AND CODES

We performed the analyses presented in this paper using the following software and codes. Codes not already published online (or provided to us by other sources) are available upon request.

SOFTWARE

R (http://www.R-project.org), MATLAB version R2008b (The MathWorks, Natick, MA, USA), SAS version 8.02, JMP version 6.0.3 (both from SAS Institute, Cary, NC, USA), BayesELO (http://remi.coulom.free.fr/Bayesian-Elo), WINSTEPS version 3.65.0 (Winsteps.com, Chicago, IL, USA), and ROC curve calculator (http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html).

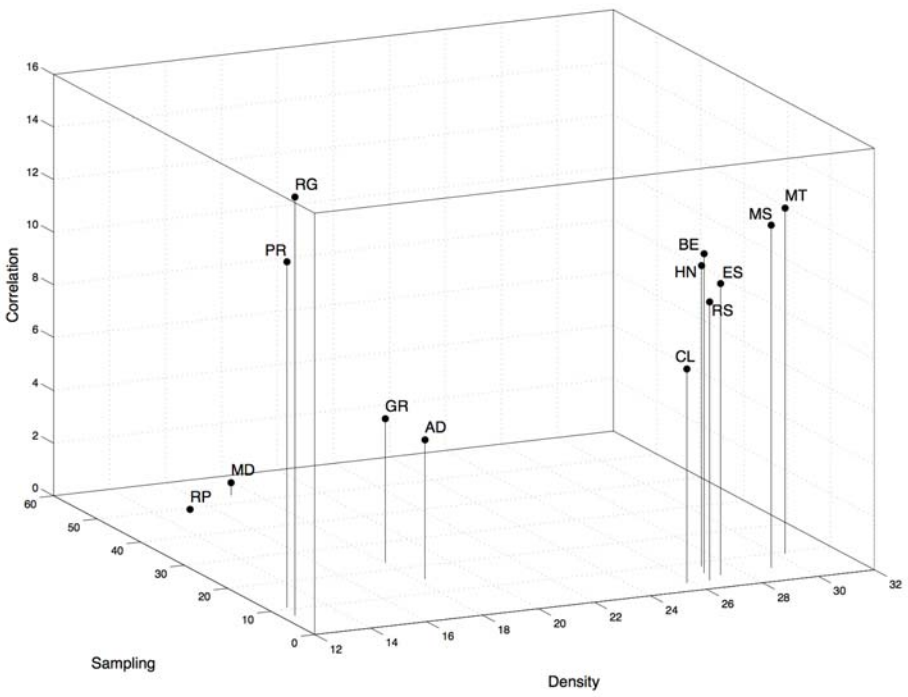
CODES

Rank products (RP) were calculated in R using the RankProd package available at http://www.bioconductor.org/packages/2.2/bioc/html/RankProd.html. The METRADISC software is available online at http://biomath.med.uth.gr, however we implemented the algorithm in MATLAB. AD was implemented in R. RG was run using SAS software with code provided to us by H.P. Piepho. ES and BE were run using Bayeselo. CL was run in MATLAB with the code "colley.m" available in Govan (2008). MS was run in R with code available online at http://www.davemease.com/football/Rcode.html. HN was run in MATLAB with the code "plackmm.m" available online at http://www.stat.psu.edu/~dhunter/code/btmatlab. GR was run in R with code provided to us by Todd Graves. RS was run with WINSTEPS. PR was run in MATLAB with the code "pagerankpow.m" available online at http://www.cs.ubc.ca/~murphyk/pmtk/doc/doc/authors/pagerankpow.m (the code was modified to be able to accept different values for *P*). MT was implemented in MATLAB from the code provided by Vasyl Pihur.

RESULTS AND DISCUSSION

When there is good agreement among ranked lists (trials) and high density of randomly distributed data-points, the difference in performance of rank-aggregating methods is small (Table 1a, 1b). Therefore, we tested the performance of 13 methods under more challenging conditions that simulate data distributions frequently observed in plant breeding trials. To measure the performance, three common measurements for rank-aggregation were used: Kendall's *tau*, Spearman's *rho*, and rank violations. Because the two correlation coefficients produced very similar results (correlation between *rho* and *tau* = 0.984) only Kendall's *tau* is shown in Table 1a and Figure 2a. Pearson's correlation coefficient between proportion of rank violations (Table 1b) and *tau* was 0.782 (with *rho* 0.749), indicating that this parameter differs somewhat from the distance-based measures. When evaluation of performance was based on proportion of rank violations, differences between methods were less pronounced and a smaller number of significant differences were detected (Table 2b). Despite variability in proportion of rank violations, all aggregate ranking methods were affected most by the sampling approach, and least by the density of data-points (Figure 2b). Differences between methods based on *tau* are discussed individually for each type of rank-aggregating method.

a)



b)

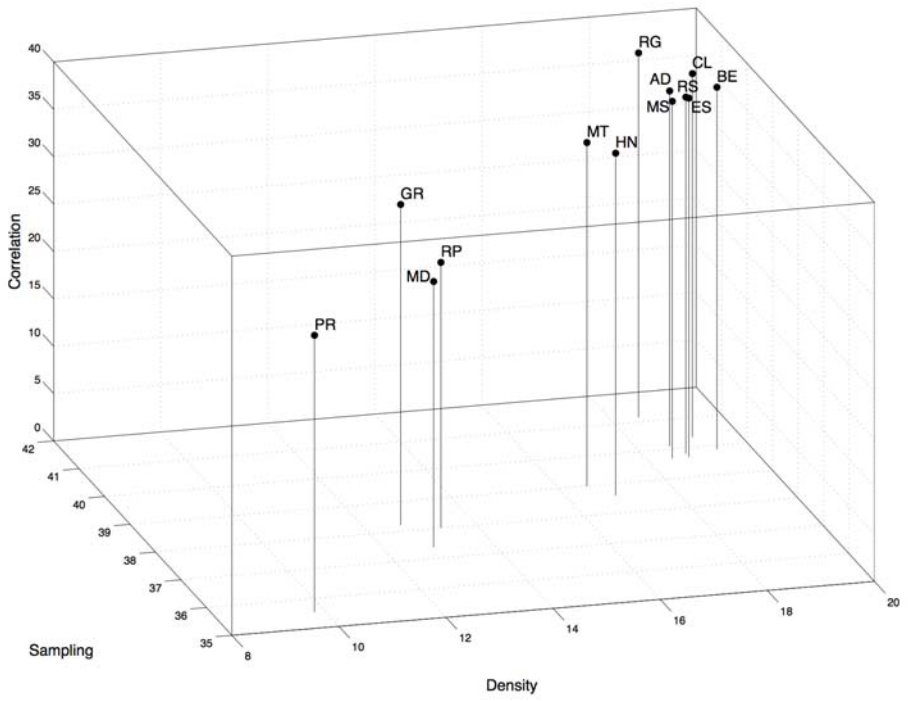


Figure 2. Effects of the three factors (density of data-points, sampling approach, and correlation between trials) on performance of 13 rank-aggregation methods. For ease of plotting, the *P*-values from three-way ANOVA were transformed into a logarithmic scale by –log(*P*). Performance of the methods was evaluated with Kendall's *tau* (2a) and proportion of rank violations (2b).

SIMPLE POSITIONAL METHODS

Simple positional methods (MD, RP) do not take into consideration which accessions are compared in each trial when aggregating ranks. Therefore, both methods significantly underperformed in the datasets where the best and worst performing accessions are never directly compared (sampling A). The factor that most significantly affects performance of these two methods (based on change in *tau*) was sampling, while the level of correlation between trials within datasets did not affect aggregate rankings. AUC for MD (0.915) and RP (0.917) were the lowest of all tested methods (Figure 3). Previously, both MD and RP have performed very well when tested on data from microarray studies (Breitling et al., 2004; Zintzaras and Ioannidis, 2008). However, unlike in our datasets, microarray-based ranks are produced from several tens of thousands of genes, of which most are overlapping in two or more microarrays. While these two methods can be used to aggregate ranks in microarray experiments, they are generally not suitable for aggregating ranks from the type of data tested in this study.

METHODS BASED ON ADJUSTED MEANS

Though the methods based on adjusted means were developed for metric data (Piepho, 2003), they performed reasonably well when tested on ranked data simulated in our study. Of the two methods, the additive model-based method performed significantly ($P < 0.001$) better overall than the regression model-based method (AUC for AD = 0.973, RG = 0.933). However, a weakness in AD was revealed when the sampling method that excludes direct comparisons between best-performing and middle-performing accessions was used to generate datasets (sampling B). The most significant factor affecting the *tau* value of AD was density of data-points, while RG was more affected by the correlation between trials in datasets. The poor performance of RG in many datasets with sparse data-points is probably due to a failure to converge (Piepho, 2003). Of these two methods, AD appears to be better suited for aggregate ranking of plant breeding trials. A further advantage of AD is that it does not require an iterative process and can be rapidly carried out on large datasets.

METHODS BASED ON PAIRED COMPARISONS

Aggregate ranking methods based on pairwise comparisons (ES, BE, CL, MS) and the Bradley-Terry model generally performed very well (AUC from 0.975 for CL to 0.978 for ES). The variability of Kendall's *tau* in all four methods was affected most by density of data-points, and least by sampling approach. A disadvantage of the Bradley-Terry model is that it does not allow estimating the likelihood under certain conditions. A problem arises, for example, when an accession performs better than any other accession in all trials in which it is tested. Then the maximum likelihood estimator for this accession is infinity and the model will produce a tie for top rank among all accessions with this estimator, regardless of which other accessions are tested in the same trials (Ford, 1957; Marden, 1995). To avoid this kind of problem several different solutions have been suggested, such as penalizing the likelihood (Mease, 2003), adding a prior (Coulom, 2008), applying the Laplace formula (Colley, 2002), or assigning a finite value (Schubert, 2000a, b). All these modifications have a relatively small effect on the final rankings, and as the number of pairwise comparisons increases the effect soon becomes negligible. Since all methods based on paired comparisons handle ties, they can be useful for combining datasets from plant breeding trials.

METHODS BASED ON MULTISTAGE MODELS

Results of the methods based on two multistage models (AUC for HN = 0.978, RS = 0.976) were comparable to those based on pairwise comparisons. The results of GR method were different from the other two multistage-based methods (AUC for GR = 0.933), since it consistently performed worse than BPM when expected top and bottom ranked accessions were never directly compared (sampling A). Variability of Kendall's *tau* for HN and RS was affected most by data-point density and least by sampling approach, while GR was affected most by sampling approach and least by correlation between trials in a dataset.

Table 1. Mean values of Kendall's *tau* (1a) and proportions of rank violations (1b) for 13 rank aggregation methods tested on 16 types of datasets.

Table 1a – Kendall's *tau*

Correlation	Density	Sampling	RP	MD	AD	RG	ES	BE	CL	MS	HN	GR	RS	PR	MT
High	Dense	R	0.94	0.94	0.94	0.93	0.95	**0.95**	0.95	0.95	0.94	0.94	0.95	0.90***	0.93*
High	Dense	A	0.40***	0.40***	0.95	0.89***	**0.96**	0.95	0.95	0.95	0.94	0.90***	0.96	0.83***	0.94
High	Dense	B	0.76***	0.92***	0.72***	0.93**	0.91***	0.95	0.89***	0.95	0.94	**0.96**	0.91***	0.84***	0.94
High	Dense	C	0.88***	0.88***	0.92	0.91	0.91	0.94	0.91	**0.94**	0.93	0.92	0.93	0.80***	0.91
High	Sparse	R	0.81***	0.83*	0.84	0.79***	0.85	0.86	0.85	0.86	0.86	**0.87**	0.85	0.74***	0.82*
High	Sparse	A	0.36***	0.37***	0.85	0.71***	**0.87**	0.86	0.80**	0.80**	0.86	0.40***	0.83	0.77***	0.83
High	Sparse	B	0.69***	0.79***	0.67***	0.79***	0.82**	**0.86**	0.80***	0.84	0.86	0.85	0.81***	0.77***	0.83*
High	Sparse	C	0.74	0.75	0.74	0.74	0.74	0.70*	0.74	**0.75**	0.69**	0.74	0.75	0.64***	0.68***
Low	Dense	R	0.91	0.90	0.90	0.81***	0.90	0.90	0.90	0.90	0.90	**0.91**	0.91	0.84***	0.89
Low	Dense	A	0.41***	0.42***	0.89	0.43***	0.90	0.89	0.89	0.89	**0.90**	0.57***	0.90	0.70***	0.88
Low	Dense	B	0.80***	0.90	0.75***	0.81***	0.89	0.90	0.89	0.90	0.90	**0.91**	0.90	0.78***	0.88*
Low	Dense	C	0.88	0.87	0.87	0.85	0.87	0.87	0.87	0.87	0.86	**0.88**	0.86	0.58***	0.83*
Low	Sparse	R	0.75	**0.76**	0.74	0.52***	0.74	0.74	0.75	0.75	0.76	0.74	0.75	0.65***	0.74
Low	Sparse	A	0.43***	0.44***	0.69	0.49***	0.70	0.69	0.65	0.64	**0.71**	0.45***	0.65	0.64***	0.67
Low	Sparse	B	0.73*	0.78	0.69***	0.54***	0.76	0.77	0.76	0.76	0.77	**0.79**	0.76	0.68***	0.75
Low	Sparse	C	0.70	**0.72**	0.71	0.68	0.71	0.70	0.71	0.71	0.69	0.70	0.70	0.62***	0.67

Table 1b – Rank Violations

Correlation	Density	Sampling	RP	MD	AD	RG	ES	BE	CL	MS	HN	GR	RS	PR	MT
High	Dense	R	0.19	0.19	0.19	0.19	**0.19**	0.19	0.19	0.19	0.19	0.19	0.19	0.19*	0.19
High	Dense	A	0.37***	0.37***	0.32	0.32	0.32	0.32	0.32	**0.32**	0.32	0.32	0.32	0.34***	0.32
High	Dense	B	0.21**	0.20	0.21*	0.20	0.20	**0.20**	0.20	0.20	0.20	0.20	0.20	0.21*	0.20
High	Dense	C	0.15	0.15	0.15	0.15	0.15	**0.15**	0.15	0.15	0.15	0.15	0.15	0.15*	0.15
High	Sparse	R	0.17	0.17	0.17	0.16	0.17	**0.16**	0.16	0.17	0.17	0.17	0.17	0.20**	0.17
High	Sparse	A	0.32***	0.33***	0.28	0.29	0.28	**0.28**	0.28	0.28	0.29	0.32***	0.28	0.33***	0.29
High	Sparse	B	0.20	0.19	0.19	0.18	0.18	**0.18**	0.18	0.18	0.18	0.18	0.18	0.20*	0.18
High	Sparse	C	0.08	0.08	0.08	0.08	0.08	**0.08**	0.08	0.08	0.08	0.08	0.08	0.11***	0.08
Low	Dense	R	0.32	0.32	0.32	0.32	**0.32**	0.32	0.32	0.32	0.32	0.32	0.32	0.33**	0.32
Low	Dense	A	0.42***	0.42***	0.40	0.43***	**0.40**	0.40	0.40	0.40	0.40	0.41***	0.40	0.43***	0.40
Low	Dense	B	0.33	0.33	0.33	0.33	0.32	**0.32**	0.32	0.32	0.33	0.33	0.32	0.33	0.33
Low	Dense	C	0.28	0.27	**0.27**	0.27	0.27	0.27	0.27	0.27	0.28	0.28	0.27	0.30***	0.28
Low	Sparse	R	0.28	0.28	0.28	0.28	0.27	**0.27**	0.27	0.27	0.28	0.28	0.27	0.30***	0.28
Low	Sparse	A	0.34	0.34	0.33	0.34	0.33	**0.33**	0.33	0.33	0.33	0.35	0.33	0.36**	0.33
Low	Sparse	B	0.29	0.29	0.28	0.29	0.28	**0.28**	0.28	0.28	0.28	0.28	0.28	0.29	0.28
Low	Sparse	C	0.18*	0.18*	0.17	**0.17**	0.17	0.17	0.17	0.17	0.17	0.18***	0.17	0.19***	0.18

The best performing method (BPM) for each type of dataset is in bold. Asterisks indicate methods that perform significantly worse than BPM at $P < 0.05$ (*), $P < 0.01$ (**), and $P < 0.001$ (***). Statistical analyses were carried out on data with more decimal places; therefore some values that appear to be identical after rounding are significant at different levels. Correlations, densities of data-points, and different samplings methods are described in Material and Methods.

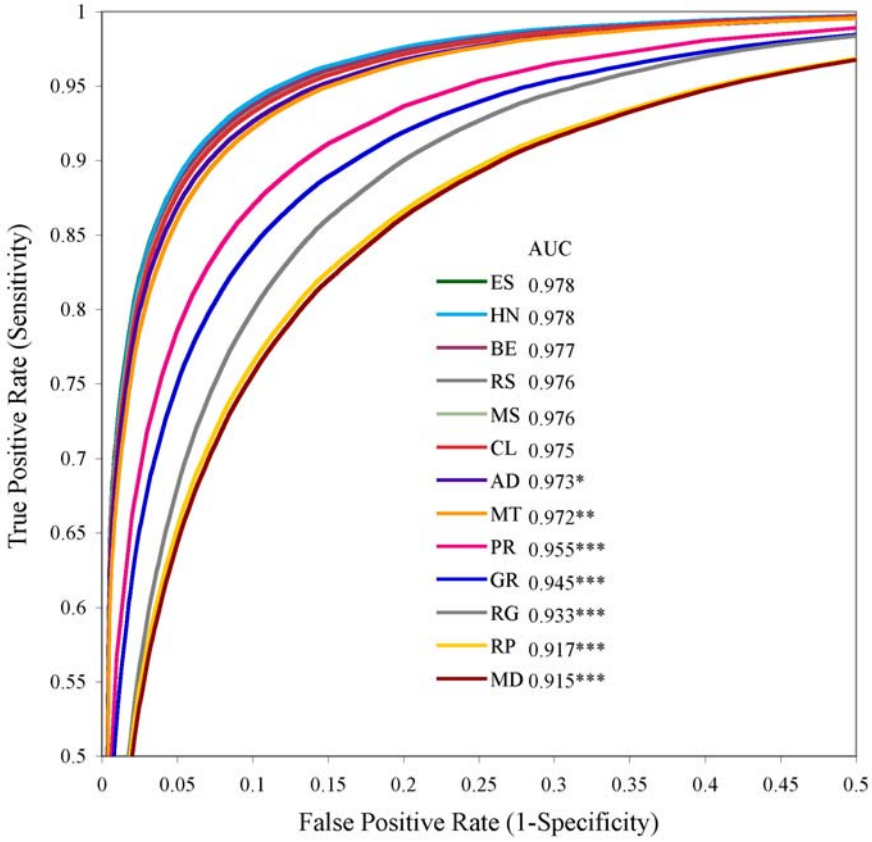


Figure 3. Receiver-operating characteristic (ROC) curve analysis carried out on 8,000 ranks from each method. The area under the ROC curve (AUC) is indicated next to the abbreviation for each method. Asterisks indicate methods that performed significantly worse than the best performing method (ES) at $P < 0.05$ (*), $P < 0.01$ (**), and $P < 0.001$ (***). For better resolution only the upper left quadrant of the whole range (from 0,0 to 1,1) is shown. Abbreviations for rank-aggregation methods are explained in Material and Methods.

Methods based on multistage models are generally used when overlapping accessions are ranked in a number of incomplete lists. Multistage models have been successfully applied in the ranking of racecar drivers (Graves et al., 2003; Hunter, 2004), racing horses (Stern, 1990), golf players (Linacre, 1992), primate intelligence (Johnson et al., 2002), crop resistance (Halekoh and Kristensen, 2008), and many areas of education and psychology (Masters, 1982). As with the methods based on Bradley-Terry, these three methods also fail to estimate maximum likelihood when an accession performs better than any other accession in all trials in which it is tested. A possible solution to this problem is the addition of a "dummy" accession to the analysis, or subtracting a fractional point value from the best and worst performing accessions (Wright, 1998). Unfortunately, neither HN nor GR can handle tied ranks, which is a significant limitation if these methods are to be considered for use in plant breeding analyses, and of the three methods tested here only RS is able to handle datasets with weak rankings (Simko and Linacre, 2010).

METHODS BASED ON MARKOV CHAINS

Aggregate rankings produced by the two methods based on MC (PR, MT) differed substantially. While *tau* values of MT were significantly ($P < 0.001$) worse than those of BPM in only one type of dataset, PR performed significantly worse than BPM in all types of datasets. Moreover, the AUC value for MT (0.972) was significantly higher than the AUC value for PR (0.955). Aggregate rankings based on PR were less affected by the three tested

factors (data-point density, sampling approach, and correlation between trials) than rankings produced by MT, which were affected more than any other method by the density of data-points. PR was originally developed to rank millions, and is currently used to rank billions, of websites of the World Wide Web (Page et al., 1998), while MT was devised to perform meta-searches by combining the results of several Web search engines to produce a collated answer (Dwork et al., 2001). Both these methods have found applications well beyond their original intentions. For example, PR has been used to rank scientific journals (Bollen et al., 2006) and papers (Chen et al., 2007), sports teams (Govan, 2008), and gene expression data in microarray experiments (Morrison et al., 2005), while MT has been applied to combine results of microarray experiments (DeConde et al., 2006) and for identification of biomarkers (Dutkowski and Gambin, 2007). Our tests of the two methods on datasets similar to those seen in plant breeding indicate that MT is more suitable for aggregating ranked data than PR. An advantage of these two MC-based methods is that they can handle large datasets more efficiently than most of the methods based on multistage or paired comparison methods.

ANALYSIS OF REAL DATA

When the earliness of tuberization data from 19 individual trials were used in marker-trait association analysis, the –log(*P*) ranged from 0.04 to 3.69. These values did not change when either ranked data (0.03 to 3.71), or 1,000 permutations of ranked data from each trial (0 to 3.58) were used as input. However, when –log(*P*) was calculated from aggregate rankings of 157 clones, the value substantially increased and ranged from 8.72 (RS) to 9.96 (PR). These results indicate that aggregating ranked data from separate trials where only a subset of accessions was tested substantially increased detection level of marker-trait association. Since –log(*P*) is a logarithmic scale, the actual change in *P* value is more than 100,000 fold (from 3.71 to 8.72). Moreover, this observed difference is not caused by more accessions being used in the analysis (157 in aggregate ranking versus 12 to 41 in individual trials) because –log(*P*) values computed from 1,000 permutations of aggregated ranks did not exceed 2.88.

SELECTING APPROPRIATE METHODS FOR AGGREGATING RANKED DATA

To select the most appropriate methods for aggregating ranks, we judged methods not only by their performance in the 16 different types of datasets, but also speed of calculation, and whether ties (weak rankings) in the original data can be handled.

Ten out of 13 tested methods performed an analysis of each dataset in less than a minute. For the other three methods (RG, MS, GR) it usually took longer to converge, sometimes 15 minutes, sometimes up to a few hours. Though the calculation time is still acceptable for the present datasets, it may become prohibitive when larger datasets are analyzed, and therefore these methods are not being considered for practical use. From the remaining methods, the consistently best aggregate rankings were observed when using ES, BE, CL, HN, and RS (AUC 0.978-0.975). Out of these five methods, HN is suitable for analysis of strong rankings only (no ties), while the remaining four methods appear to be well suited for combining ranked lists from plant breeding trials. All four of these methods show high AUC and *tau* values, and low proportions of rank violations.

In addition to rankings, we performed a limited test of ratings produced by the four methods on three additional datasets (data not shown). These datasets were built with 100 individuals and 20 trials, and no data points were removed through sampling. Observed values were generated using a normal, uniform, and bimodal distribution, respectively. The Pearson's correlation coefficient between the means of the original data and the final aggregate rating was above 0.97 for all methods and distributions, with the exception of RS where coefficient was 0.92 for normal distribution. This indicates that not only rankings, but also ratings produced by these four methods strongly correlate with the expected data distribution.

However, when selecting the best methods we did not explicitly consider other factors, such as capability to combine trials or studies with different levels of reliability, or calculating other statistics such as standard deviation of aggregate rankings. Also, much larger datasets with high densities of data-points could lead instead to selection of a method that is even less computationally demanding.

There are several other methods that can be used to produce aggregate rankings from partially ranked data. These include a linear programming approach (implemented in MinV) (Coleman, 2005) and a branch-and-bound search (Cook et al., 2007), which determine optimal rankings based on minimizing the number of rank violations, and RankAggreg (Pihur et al., 2008) and TopKCEMC (Lin and Ding, 2009), which use Kendall's *tau* and/or Spearman's *rho* to find aggregate rankings that most closely match the partial rankings. However, all these methods (except MinV) use strong rankings only, meaning no ties can be input, and are very computationally demanding – much more so than any method tested in this study.

CONCLUSIONS

Approaches for combining metric data from multiple series of plant breeding trials have been extensively studied (e.g. Piepho, 2003), but little information is available regarding ranked data when only a subset of accessions is evaluated in each trial. To test the performance of different aggregate-ranking methods, we designed and generated several types of datasets that resemble those typically found in plant breeding trials. Our analysis indicates that three methods based on Bradley-Terry (ES, BE, CL) and one based on Rasch (RS) models performed better than the other tested methods when factors such as fitness of aggregate rankings, time required for analyses, and the ability to analyze weak rankings were considered. Accuracy of these rank-aggregating methods improved with an increased density of data-points.

We further showed on real data that combining ranked lists from several trials significantly improves the power to detect factors of interest despite noise in the datasets. In our example the significance of linkage between a marker and a trait of interest was substantially increased (as indicated by a drop in *P*-value by a factor of 100,000) when tests were performed on aggregated ranks as compared to data from the original individual trials. The ability to combine heterogeneous data from independent trials has important ramifications for data analysis. Results from our study indicate that this kind of meta-analysis is more powerful than individual analyses. Application of the aggregate ranking approach is not limited to plant breeding trials, but can be applied also in other areas of agricultural and biological research with similar distributions of data. For example, in genetics several datasets containing phenotypic information can be combined into one, which can then be used in association mapping studies

ACKNOWLEDGEMENTS

The authors would like to thank Rémi Coulom, Tal Raviv and Vasyl Pihur for valuable discussions, code, and examples.

REFERENCES

Bollen, J., Rodriguez, M.A., Van de Sompel, H. (2006). Journal status. *Scientometrics* 69, 669–689.

Bradley, R.A., Terry, M.E. (1952). Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika* 39, 324–345.

Breitling, R., Armengauda, P., Amtmanna, A., Herzyk, P. (2004). Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters* 573, 83–92.

Chen, P., Xie, H., Maslov, S., Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics* 1, 8–15.

Coleman, J.B. (2005). Minimizing game score violations in college football rankings. *Interfaces* 35, 483–496.

Colley, W.N. (2002). Colley's bias free college football ranking method: The Colley matrix explained. http://wwwcolleyrankingscom/matratepdf.

Cook, W.D., Golany, B., Penn, M., Raviv, T. (2007). Creating a consensus ranking of proposals from reviewers' partial ordinal rankings. *Computers & Operations Research* 34, 954–965.

Coulom, R. (2008). Whole-history rating: A Bayesian rating system for players of time-varying strength. In: van der Herik, H.J., Xu, X., Ma, Z., Winands, M.H.M. (Eds.) *Computer and Games, 6th International Conference*, Beijing, China, September 29 – October 1, 2008.

DeConde, R.P., Hawley, S., Falcon, S., Clegg, N., Knudsen, B., Etzioni, R. (2006). Combining results of microarray experiments: A rank aggregation approach. *Statistical Applications in Genetics and Molecular Biology* 5, Art. 15.

Dutkowski, J., Gambin, A. (2007). On consensus biomarker selection. *BMC Bioinformatics* 8, S5.

Dwork, C., Kumar, R., Naor, M., Sivakumar, D. (2001). Rank aggregation for the web. *Proceeding of the 10th International conference on World Wide Web, WWW10*, Hong Kong, May 1 – 5, 2001, 613–622.

Ford, L.R.J. (1957). Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly* 64, 28–33.

Govan, A.Y. (2008). *Ranking theory with application to popular sports*. Ph.D. Thesis, North Carolina State University, Raleigh, NC.

Graves, T., Reese, C.S., Fitzgerald, M. (2003). Hierarchical models for permutations. *Journal of the American Statistical Association* 98, 282–291.

Halekoh, U., Kristensen, K. (2008). Evaluation of treatment effects by ranking. *Journal of Agricultural Science* 146, 471–481.

Hanley, J.A., McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 143, 29–36.

Hartung, K., Piepho, H.P. (2005). A threshold model for multiyear genebank data based on different rating scales. *Crop Science* 45, 1045–1051.

Hsu, J. (1981). Simultaneous confidence intervals for all distances from the 'best'. *Annals of Statistics* 9, 1026–1034.

Hunter, D.R. (2004). MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics* 32, 384–406.

Johnson, V.E., Deaner, R.O., van Schaik, C.P. (2002). Bayesian analysis of rank data with application to primate intelligence experiments. *Journal of the American Statistical Association* 97, 8–17.

Lin, S., Ding, J. (2009). Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA studies. *Biometrics* 65, 9–18.

Linacre, J.M. (1992). Objective measurement of rank-ordered objects. In: Wilson, M. (Ed.) *Objective Measurement: Theory into Practice*, Ablex, Norwood, NJ, 195–209.

Luce, R.D. (1959). *Individual choice behavior: A theoretical analysis*. Wiley, New York, NY.

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

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika* 47, 149–174.

Mease, D. (2003). A penalized maximum likelihood approach for the ranking of college football teams independent of victory margins. *The American Statistician* 57, 241–248.

Morrison, J.L., Breitling, R., Higham, D.J., Gilbert, D.R. (2005). GeneRank: Using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics* 6, 233.

Norris, J.R. (1997). *Markov chains*. Cambridge University Press, Cambridge.

Page, L., Brin, S., Motwani, R., Winograd, T. (1998). *The PageRank citation ranking: Bringing order to the web*. Technical report, Computer Science Department, Stanford University, Stanford, CA, USA

Piepho, H.P. (2003). Model-based mean adjustment in quantitative germplasm evaluation data. *Genetic Resources and Crop Evolution* 50, 281–290.

Pihur, V., Datta, S., Datta, S. (2008). Finding common genes in multiple cancer types through meta–analysis of microarray experiments: A rank aggregation approach. *Genomics* 92, 400–403.

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

Schubert, F, (2000a). Chess statistics - Rating estimation of chess programs as a mathematical problem (part 1) (in German). *Computerschach und Spiele* 2000, 29–34.

Schubert, F. (2000b). Chess statistics - Rating estimation of chess programs as a mathematical problem (part 2) (in German). *Computerschach und Spiele* 2000, 45–50.

Simko, I., Linacre, J.M. (2010). Combining partially ranked data in plant breeding and biology: II. Analysis with Rasch model. *Communications in Biometry and Crop Science* 5, 56–65.

Simko, I., Vreugdenhil, D., Jung, C.S., May, G.D. (1999). Similarity of QTLs detected for in vitro and greenhouse development of potato plants. *Molecular Breeding* 5, 417–428.

Stern, H. (1990). Models for distribution on permutations. *Journal of the American Statistical Association* 85, 558–564.

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review* 34, 273–286.

Thurstone, L.L. (1931). Rank order as a psychophysical method. *Journal of Experimental Psychology* 14, 187–201.

van den Berg, J.H., Ewing, E.E., Plaisted, R.L, McMurry, S., Bonierbale, M.W. (1996). QTL analysis of potato tuberization. *Theoretical and Applied Genetics* 93, 307–316.

Wright, B.D. (1998). Estimating measures for extreme scores. *Rasch Measurement Transactions* 12, 632–633.

Zintzaras, E., Ioannidis, J.P.A. (2008). Meta-analysis for ranked discovery datasets: Theoretical framework and empirical demonstration for microarrays. *Computational Biology and Chemistry* 32, 39–47.