Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2019, Vol. 28(8) 2292–2304

DOI: 10.1177/0962280218768326
journals.sagepub.com/home/smm

An omnibus test for the global null hypothesis

Andreas Futschik,[1,2,3] Thomas Taus[3,4] and Sonja Zehetmayer[5]

Abstract

Global hypothesis tests are a useful tool in the context of clinical trials, genetic studies, or meta-analyses, when researchers are not interested in testing individual hypotheses, but in testing whether none of the hypotheses is false. There are several possibilities how to test the global null hypothesis when the individual null hypotheses are independent. If it is assumed that many of the individual null hypotheses are false, combination tests have been recommended to maximize power. If, however, it is assumed that only one or a few null hypotheses are false, global tests based on individual test statistics are more powerful (e.g. Bonferroni or Simes test). However, usually there is no a priori knowledge on the number of false individual null hypotheses. We therefore propose an omnibus test based on cumulative sums of the transformed p-values. We show that this test yields an impressive overall performance. The proposed method is implemented in an R-package called *omnibus*.

Keywords

Multiple testing, global null hypothesis, omnibus test, meta-analysis, experimental evolution

1 Introduction

When testing multiple hypotheses, the global null hypothesis is often of specific interest. It states that none of the individual null hypotheses is false. In some applications, rejecting the global null can be a goal in itself, whereas in other situations such a test may occur as part of a more sophisticated multiple test procedure. Think for instance of the closure test principle, where the global null needs to be rejected before looking at specific tests. Also, in an ANOVA, the global null is usually tested before testing for pairwise differences.

In meta-analysis, rejecting the global null implies an effect at least under some circumstances. Another application is experimental evolution, where several replicate populations of micro- or higher organisms are maintained under controlled laboratory conditions and their response to selection pressures is studied. Further applications, where such a test is of interest in its own merit, are testing for overall genomic differences in gene expression and signal detection.[1,2]

Several approaches to test the global null hypothesis are known. If we assume alternative scenarios where all or most null hypotheses do not hold, combination tests (e.g. Fisher's combination test[3] or Stouffer's z test[4]), that sum up two or more independent transformed p-values to a single test statistic, have been recommended to maximize power. If, however, it is assumed that the null hypothesis holds in most cases, global tests based on individual test statistics are more powerful (e.g. Bonferroni, Simes).[5] If a larger number of hypotheses are tested, and the alternative hypothesis holds sufficiently often, goodness of fit tests for a uniform distribution of p-values could also be used. They test, however, for any type of deviation from uniformity and do not focus specifically on too

[1]Department of Applied Statistics, Johannes Kepler University, Linz, Austria
[2]Kavli Institute for Theoretical Physics, UC Santa Barbara, Santa Barbara, USA
[3]Vienna Graduate School of Population Genetics, University of Veterinary Medicine, Vienna, Austria
[4]Institute of Population Genetics, University of Veterinary Medicine, Vienna, Austria
[5]Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria

Corresponding author:
Sonja Zehetmayer, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, Vienna 1090, Austria.
Email: sonja.zehetmayer@meduniwien.ac.at

small p-values. Under more specific models, such as the comparison of several normal means, more specialized tests such as a Tukey's multiple range test or an ANOVA are further options.

Higher criticism (HC) and checking for *overall significance* are alternative terms used instead of global testing. Originating from biblical science, the term *higher criticism* was first used by Tukey[6] in a statistical context. Making the point that a certain number of falsely rejected null hypotheses can be expected when testing several null hypotheses at level α, he then proposed a *second level significance test* to check for overall significance. Later, Donoho and Jin[7,8] provided an asymptotic analysis of this and related tests, when the number of hypotheses tends to infinity. Their results show that there are situations where there is sufficient power to detect deviations from the global null hypothesis, but no chance to reliably identify in which cases the alternative holds.

Our focus is on a general situation where independent p-values are available from several hypothesis tests that are assumed to be uniformly distributed under the null hypothesis. As there is often no a priori knowledge on the number of false individual null hypotheses, we propose a test that enjoys good power properties, both if few and many null hypotheses are false. Our test is based on cumulative sums of the (possibly transformed) sorted p-values.

In comparison to other available methods, our simulations show that this test yields an excellent overall behavior. It typically performs better than combination tests, if the alternative holds in only a few cases. If the alternative holds in most cases, it performs better than the Bonferroni and Simes test. The performance relative to methods that combine evidence across all p-values tends to be even better under those one-sided testing scenarios, where parameters are in the interior of the null hypothesis for some of the tests. For these tests, the corresponding p-values will be stochastically larger than uniformly distributed ones, reducing in particular the power of combination tests.

We also present real data applications in the context of meta-analysis and experimental evolution.

2 Testing the global null hypothesis based on p-values

Consider a multiple testing procedure with m null hypotheses H_{0i}, $i = 1, \ldots, m$, of which m_0 are true and m_1 are false. We assume that m, possibly different, hypothesis tests are carried out leading to stochastically independent p-values $p_1, \ldots, p_m$. Our focus is on testing the global null

$$H_0 = \bigcap_{i=1}^{m} H_{0i}$$

i.e. that none of the null hypotheses is false. We assume that the p-values are either uniformly distributed

$$p_i \sim U_{[0,1]}$$

under the global null hypothesis H_0 or that the p-values are stochastically larger than uniformly distributed ones. In other words, we assume that $P(p_i \leq x) \leq x$ for $0 \leq x \leq 1$.

Some tests for the global null hypothesis use a combined endpoint, summing up the evidence across all available p-values to a single test statistic, e.g. Fisher's combination function[3] or Stouffer's test.[4] Alternatively, other approaches focus on those individual test statistics that lead to extreme p-values, such as in the Bonferroni and Simes tests. As combination tests aggregate evidence across all hypotheses, these tests are particularly powerful when there are (small) effects in many considered null hypotheses. When there are only a few (large) effects, global tests based on individual test statistics are more powerful. Other approaches are goodness of fit tests or HC.

2.1 Omnibus test

2.1.1 General outline

Starting with independent p-values $p_1, \ldots, p_m$, we denote the sorted p-values by

$$p_{(1)} \leq \cdots \leq p_{(m)}$$

and transform them with a monotonously decreasing function $h(\cdot)$ so that small p-values lead to large scores. Possible choices for $h(\cdot)$ will be discussed below. Next, we obtain the L-statistics $S_i = \sum_{j=1}^{i} h(p_{(j)})$, $i = 1, \ldots, m$. Each of these partial sums could in principle be chosen as a test statistic for the global test and the best choice in terms of power for a specific scenario will depend both on (m_0, m_1) and the respective effect sizes. Since these

quantities are unknown, we propose to select the most unusual test statistic out of S_i, $1 \le i \le m$. If the scores S_i were approximately normally distributed, we could standardize them to figure out how unusual they are. Here, however, the distribution of the S_i with small index will be closer to an extreme value distribution. Therefore, we transform the sums using the distribution function G_i of S_i under the global null hypothesis, and take

$$T^* = \max_{1 \le i \le m} G_i(S_i)$$

as our test statistic. Although the cumulative sums S_i may be viewed as L-statistics, and conditions that ensure the asymptotic normality of L-statistics $m \to \infty$ are known,[9] these conditions are not satisfied for some of the S_i, and furthermore the number of hypotheses is small to moderate. We therefore estimate the distribution of T^* by simulating uniformly distributed p-values under the global null. Note, however, that some of our considered transformations lead to situations where exact null distributions are available for S_i. This turns out to be the case in particular when the transformed p-values $h(p_j)$ follow a uniform, exponential, or (skewed) normal distribution.[10,11]

Later on, we will consider four transformations $h(p)$ in more detail:

- $h(p) = 1 - p$ (omnibus p)
- $h(p) = -\log p$ (omnibus *log* p)
- $h(p) = z_{1-p}$ (omnibus z)
- $h(p) = p^{-\alpha}$ with $\alpha = 0.5$ (omnibus power)

Here, z_{1-p} denotes the $1–p$ quantile of the standard normal distribution. In principle, any monotonously decreasing function can be applied as transformation. The above four transformations were selected to give different weight to extreme p-values. Note that for small enough p, we have that

$$1 - p \le z_{1-p} \approx \sqrt{2\log(1/p)} \le \log(1/p) \le p^{-\alpha}$$

implying that a few very small p-values are most influential when using $h(p) = p^{-\alpha}$. Based on our simulation results, we will discuss the choice of $h(\cdot)$ in more detail. Here, we only mention that $h(p) = -\log(p)$ leads to a test with a particularly good worst case performance.

2.2 Alternative test statistics

We briefly explain the most popular approaches that use p-values for testing the global null hypothesis.

2.2.1 Fisher's combination test

Fisher[3] proposed the combined test statistic given by $T = -\sum_{i=1}^{m} 2\log p_i$. Under the assumption of independent uniformly distributed p-values, the null distribution is $T \sim \chi^2_{2m}$.

2.2.2 Stouffer's z test

Based on z-values $Z_i = z_{1-p_i}$, the combined test statistic is given by $Z = \sum_{i=1}^{m} Z_i/\sqrt{m}$. Assuming again independent uniformly distributed p-values under the global null, it can be easily seen that $Z \sim N(0,1)$.[4]

2.2.3 Bonferroni test

The Bonferroni test rejects the global null hypothesis, if the minimum p-value falls below α/m, i.e. $\min_i p_i \le \alpha/m$ (see, e.g. Dickhaus[12]). The Bonferroni test controls the family-wise error rate at level α in the strong sense. The test makes no assumption on the dependence structure of endpoints. For independent test statistics, α/m may be replaced by the slightly more liberal upper bound $1 - (1-\alpha)^{1/m}$.

2.2.4 Simes test

An improvement of the Bonferroni test in terms of power was proposed by Simes.[5] For the m hypotheses H_{0i}, $i = 1, \ldots, m$, with p-values p_i, the Simes test rejects the global null hypothesis if for some $k = 1, \ldots, m$, $p_{(k)} \le \alpha k/m$. In the last decades, the Simes test has become very popular for testing individual hypotheses controlling the False Discovery Rate.[13]

2.2.5 Higher criticism

Based on an idea by Tukey,[6] Donoho and Jin[7,8] introduced the HC to test the global null hypothesis of no effect for independent hypotheses. It is defined by

$$\mathrm{HC}_m^* = \max_{1 \le i \le \alpha_0 m} \left\{ \sqrt{m} \frac{i/m - p_{(i)}}{\sqrt{p_{(i)}(1 - p_{(i)})}} \right\}$$

α_0 is a tuning parameter often set to 1/2 and has been studied in particular for large-scale testing problems.

2.2.6 Goodness of fit tests

For our global test problem of independent p-values and under a point null hypothesis, the p-values p_i, $i = 1, \ldots, m$, often follow a uniform distribution $U(0, 1)$. Thus, any goodness of fit test for uniformity, such as the Kolmogorov–Smirnov (KS), the Chi-square, and the Cramer–von Mises tests also provide tests for the global null hypothesis. The KS test, for example, uses the maximum distance between the empirical distribution function of the observed p-values and the uniform distribution function, $D_n = sup_{0 \le x \le 1} |F_n(x) - x|$ as test statistic. A disadvantage of goodness of fit tests in our context is that they test not only for smaller than expected p-values but against any deviation from uniformity. As also confirmed by our simulations, these tests therefore provide lower power compared to more specialized tests in our situation (data not shown).

3 Results

We start our simulation study by comparing the power, i.e. 1–p(type II error), of our test when different transformations $h(\cdot)$ are used. It will turn out that $h(p) = -\log p$ leads to particularly good overall behavior across different scenarios, and we thus focus on this transformation when comparing our approach with alternative tests for the global null, such as the Bonferroni and the Simes procedure, as well as Fisher's and Stouffer's combination tests. Although typically used for a large number of hypotheses, we will also consider HC as a competing method (with the tuning parameter $\alpha_0 = 0.5$). As the asymptotic approximations do not necessarily hold for small numbers of hypotheses, we simulate critical values under the null model for this test.

We simulate different scenarios by varying both the total number m of hypotheses, and the number m_1 of instances where the alternative holds. We assume independence between the p-values, which was a condition in our derivation of the omnibus test.

Although our test is based on p-values that may arise in a multitude of settings, we want to specify effect sizes and alternative distributions in an intuitive way, and therefore compute our p-values from normally distributed data with known variance $\sigma^2 = 1$ and equal sample sizes n. More specifically, we consider the one-sample z-test for one-sided hypotheses

$$H_{0i} : \mu_i = 0 \quad \text{versus} \quad H_{1i} : \mu_i > 0, \; i = 1, \ldots, m$$

for the mean of the observations.

In the simulations, we first assume that all alternatives have the same mean effect Δ/σ and for the true null hypotheses $\Delta = 0$. Later on, we also consider the following setups:

(i) Negative effect sizes that are in the interior of the null hypotheses: We assume that under the true null hypothesis, the data have a negative effect size of $-\Delta/\sigma$ and under the alternative hypothesis a positive effect size of Δ/σ.
(ii) Different effect sizes of alternative hypotheses: We assume randomly chosen exponentially distributed effect sizes with a rate parameter of $3\sqrt{m_1}$.
(iii) Different effect sizes of alternative hypotheses and different effect sizes in the interior of the null hypotheses: We assume randomly chosen exponentially distributed effect sizes with a rate parameter of $3\sqrt{m_1}$ or $-3\sqrt{m_1}$, respectively.

All computations were performed using the statistical language R,[14] the Fisher's and the Stouffer's combination tests were calculated using the function combine.test in the survcomp package.[15] Our proposed omnibus method and the HC are implemented in the R-package omnibus available at https://github.com/ThomasTaus/omnibus.

For each scenario at least 10,000 simulation runs were performed.

For all following simulation results, the methods control the type I error at 5% if the global null hypothesis is true (simulation results can be found in the online supplemental material).

3.1 Influence of the chosen transformation on the omnibus method

Figure 1 shows power curves for the omnibus test using the four proposed transformations. We consider $m = 10$, $m_1 \in \{1, 3, 5, 10\}$, and $\Delta/\sigma = 0.3/\sqrt{m_1}$. These variants show similar power values for a lot of scenarios. Nevertheless, the performance of the power ("power") and identity transforms ("p") seems to be somewhat less satisfactory. In particular, the power transform performs considerably worse when the alternative is true in several instances, while giving only slightly better results in the case of only one true alternative. The z and *log p* transforms both show a good overall behavior. The *log p* transform performs slightly better for small m_1 (i.e. a few larger effects), whereas the omnibus z method turns out to be slightly better if m_1 is large (i.e. several smaller effects). Similar results were observed for $m = 5$ and $m = 20$ (see Figures 1 and 2 in the online supplemental material). We provide a between methods comparison of the worst case power across all possible choices of m_1 with constant cumulative effect sizes. According to Table 1, the omnibus *log p* transform slightly outperforms the z transform. Thus, we will use the *log p* transformation with our omnibus test subsequently.

3.2 Power comparison between different testing methods

Figure 2 shows power curves for omnibus *log p*, Bonferroni test, Simes test, Fisher's combination test, Stouffer's z test, and HC for $m = 10$, $m_1 \in \{1, 3, 5, 10\}$, and $\Delta/\sigma = 0.3/\sqrt{m_1}$. It can be seen that the omnibus method is among the top methods concerning power for all scenarios (black solid curves).

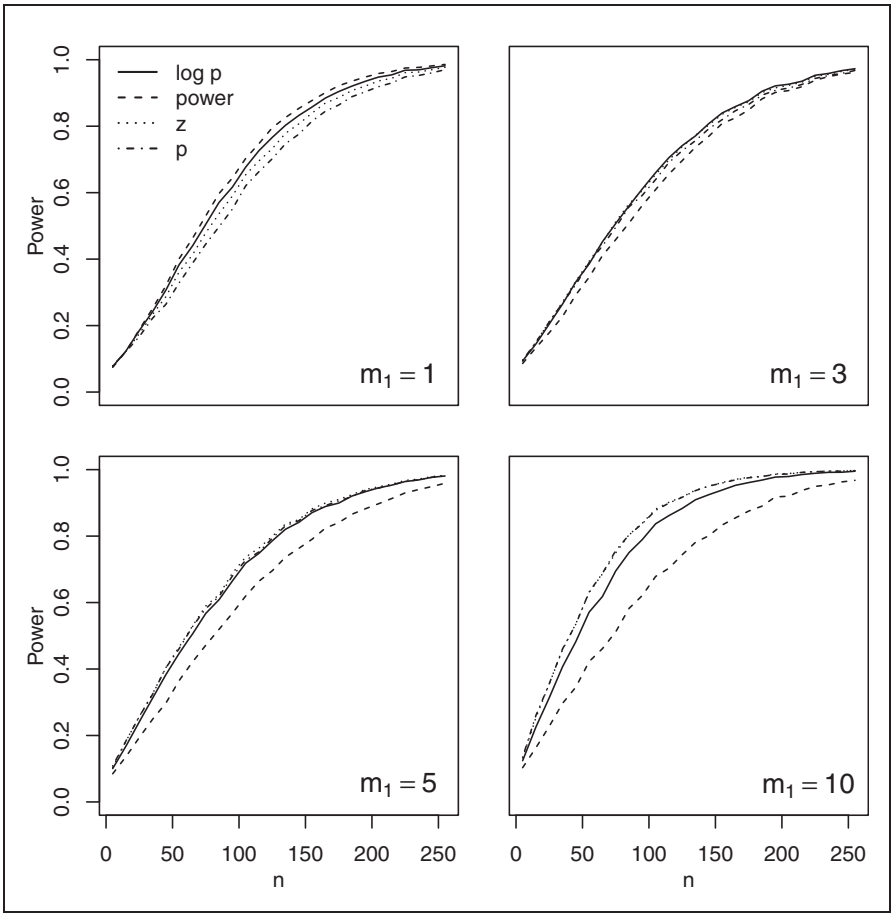

Figure 1. Power values for omnibus *log p*, *power*, *z*, and *p* are given for increasing *n*, $m = 10$, $m_1 \in \{1, 3, 5, 10\}$, $\Delta/\sigma = 0.3/\sqrt{m_1}$.

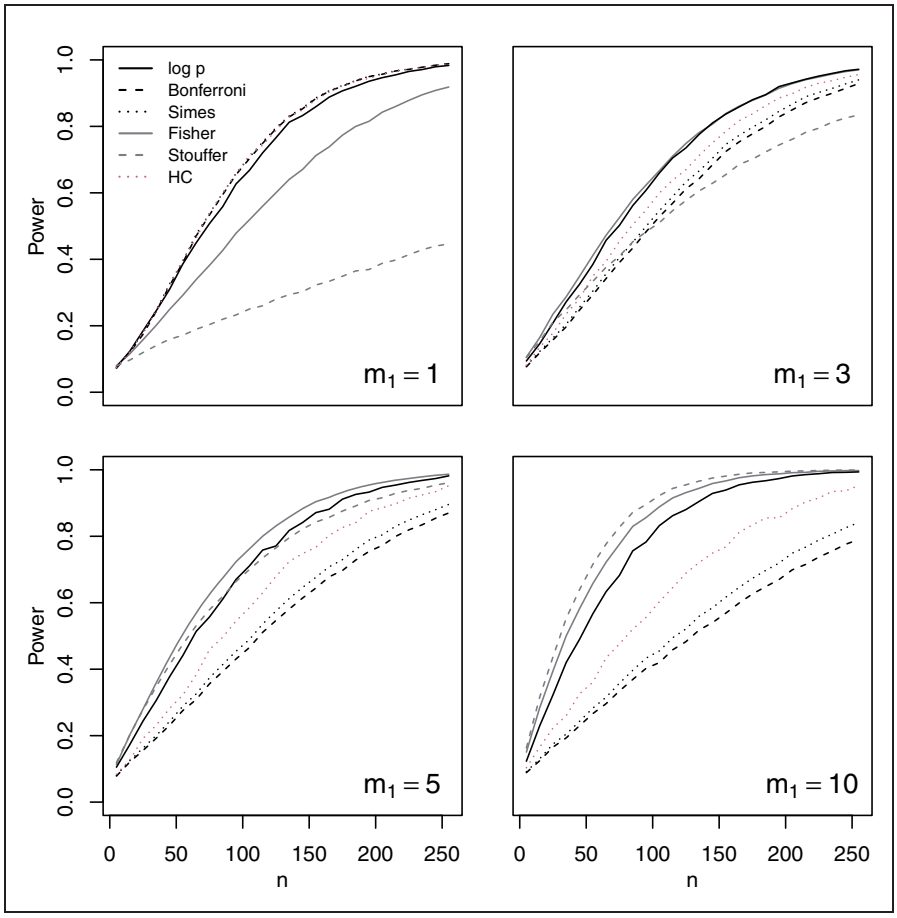

Figure 2. Power values for increasing n, $m = 10$, $m_1 \in \{1, 3, 5, 10\}$, $\Delta/\sigma = 0.3/\sqrt{m_1}$ for omnibus *log p*, Bonferroni, Simes, Fisher's test, Stouffer's test, and HC.

Table 1. Minimax power.

	$m = 10$		$m = 20$		$m = 1000$	
	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
Omnibus *log p*	0.63	0.92	0.50	0.84	0.23	0.50
Omnibus *z*	0.62	0.92	0.49	0.83	0.23	0.49
Omnibus *p*	0.59	0.90	0.46	0.82	0.22	0.48
Bonferroni	0.41	0.68	0.28	0.47	0.11	0.18
Simes	0.44	0.73	0.30	0.52	0.12	0.19
Fisher	0.49	0.83	0.35	0.67	0.15	0.28
Stouffer	0.24	0.39	0.16	0.25	0.09	0.11
HC half	0.53	0.86	0.40	0.73	0.14	0.30

Note: Worst case power values for m_1 from 1 to m (minimum over all simulation scenarios) for $n = \{100, 200\}$, $m = \{10, 20, 1000\}$, $\Delta/\sigma = 0.3/\sqrt{m_1}$.

The Bonferroni and Simes methods give the best power results in the case of only one false null hypothesis, $m_1 = 1$; however, the difference to the omnibus *log p* variant of our test is only marginal. For increasing m_1, the power of the Bonferroni and Simes methods is inferior compared to all other methods. As expected, the Simes test outperforms the Bonferroni procedure (or is equal), though, for the considered scenarios the improvement in power is only small.

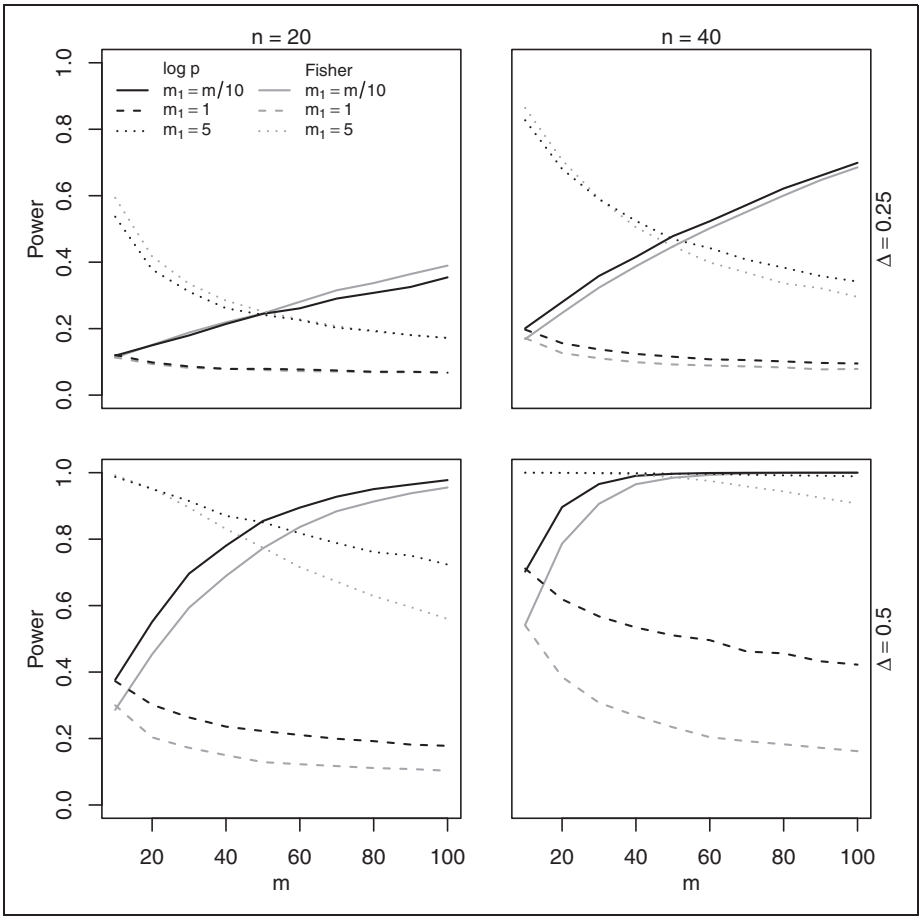

Figure 3. Omnibus *log p* (black line) and the Fisher's combination test (gray line) for $n \in \{20, 40\}$, $\Delta \in \{0.25, 0.5\}$, increasing m and $m_1 = m/10$, $m_1 = 1$, or $m_1 = 5$, respectively.

The Fisher's combination test is slightly superior in scenarios with large m_1 in comparison to the omnibus tests; however, it has low power for small m_1, e.g. for scenarios with $m_1 = 1$ the omnibus test has nearly 20 percentage points higher power than the Fisher's test. The Stouffer's test only shows competitive power values for high number of false null hypotheses for the considered scenarios. In contrast, the HC method for $\alpha_0 = 0.5$ has similar power values as Bonferroni and Simes for $m_1 = 1$, for increasing m_1 the omnibus *log p*, Fisher's, and Stouffer's tests are clearly more powerful. Note that similar results were observed for $m = 5$ and $m = 20$ (see Figures 3 and 4 in the online supplemental material).

3.2.1 Worst case behavior

We assess also the overall behavior of the statistical tests we considered by looking at the minimax power across scenarios that involve all possible numbers m_1 of true alternative hypotheses. We define the minimax power as the lowest power across all these scenarios. With m_1 alternatives, the individual effect size was chosen $\Delta/\sigma = \gamma/\sqrt{m_1}$. This leads to a constant cumulative effect size of $\frac{\sqrt{m_1}\gamma}{\sqrt{m_1\sigma^2}} = \gamma/\sigma^2$. This constant cumulative effect size would also lead to equal power for any value of m_1 under a simplified scenario where a likelihood ratio test of the global null versus m_1 false null hypotheses with known index under the alternative is applied. In the theoretical case that both m_1 and the position of the m_1 hypotheses are known, an optimal test could be obtained this way that leads to constant non-centrality parameters for all values of m_1. Table 1 uses $\gamma = 0.3$, leading to intermediate power values. As can be seen, the considered omnibus tests outperform the other tests with respect to the worst case behavior, with the omnibus *log p* test performing best in this sense.

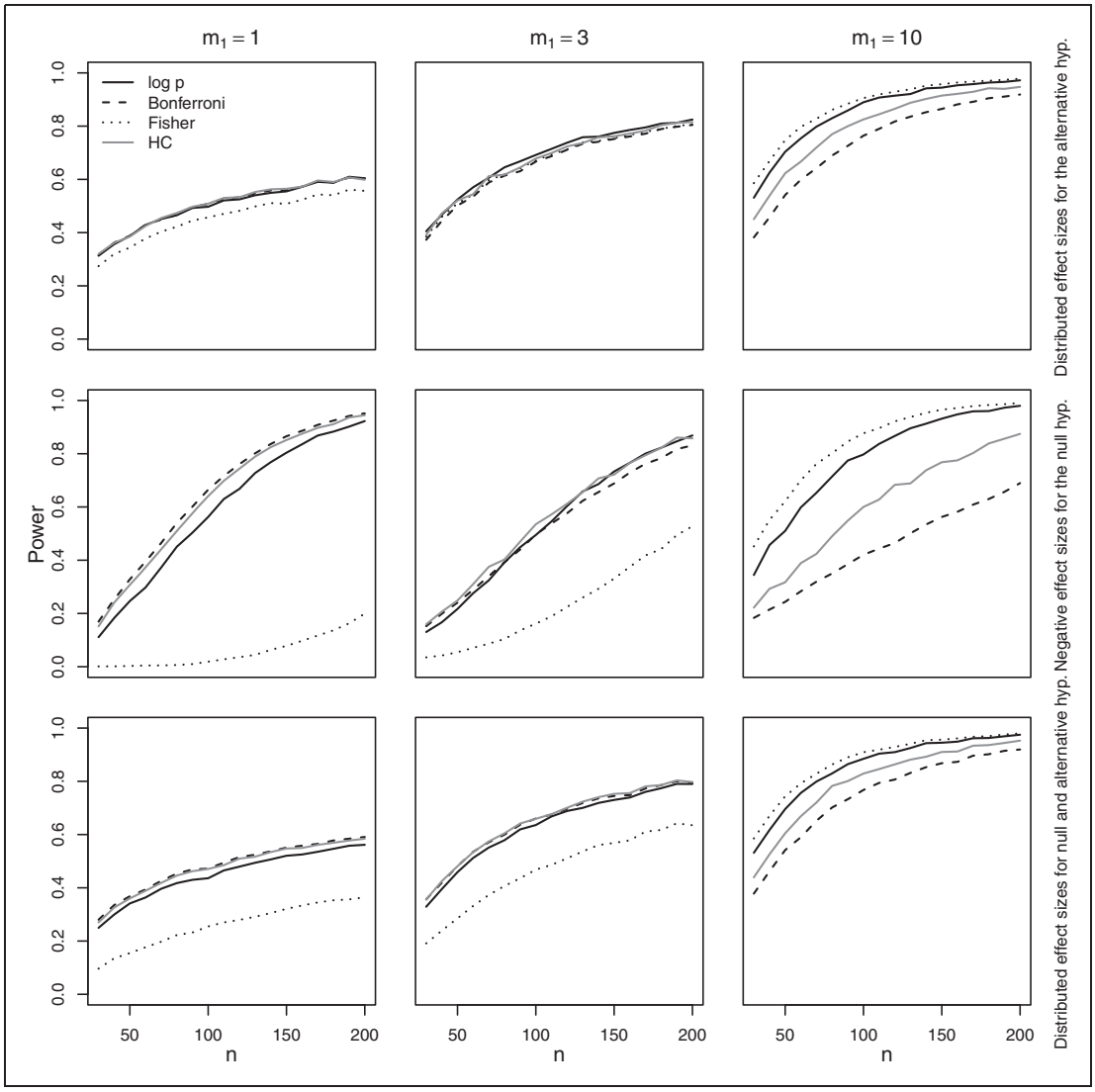

Figure 4. Power values are given for increasing n, $m = 10$, $m_1 \in \{1, 3, 10\}$ for omnibus *log p*, Bonferroni test, Fisher's combination test, and HC. The first row shows results for distributed effect sizes of alternative hypotheses according to an exponential distribution with rate parameter $3\sqrt{m_1}$. The second row shows results for $\Delta/\sigma = -0.3/\sqrt{m_1}$ under the null hypothesis and $\Delta/\sigma = 0.3/\sqrt{m_1}$ under the alternative. The third row shows results for distributed effect sizes according to an exponential distribution under the alternative as well as under the null hypothesis.

3.2.2 *Behavior for small numbers m_1 of true alternatives*

To further compare the power of our omnibus *log p* test and Fisher's combination test, we performed simulations when m_1 is small, either in absolute terms or compared to m. More specifically, we considered $m_1 = 1$, $m_1 = 5$, as well as $m_1 = m/10$. We assigned the same fixed effect sizes $\Delta/\sigma \in \{0.25, 0.5\}$ to each alternative hypothesis. Figure 3 shows the power curves of the omnibus test (black curves) and the Fisher's combination test (gray curves) for $n \in \{20, 40\}$, and increasing m. The omnibus test provides a higher power in most scenarios. Only in the situation of small effect sizes ($\Delta = 0.25$), the Fisher's combination test behaves better under some circumstances. This occurs in particular when $m_1 = 5$, and m fairly small, implying a fairly large proportion m_1/m of alternatives. Note, however, that the difference in power is small in these cases compared to the excess power of the omnibus test for larger effect sizes.

3.3 Distributed/negative effect sizes

In Figure 4 (first row), we show simulation results for distributed effect sizes with a mean effect Δ distributed according to an exponential distribution with a rate parameter of $3\sqrt{m_1}$ for $m_1 \in \{1, 3, 10\}$, $m = 10$. Generally,

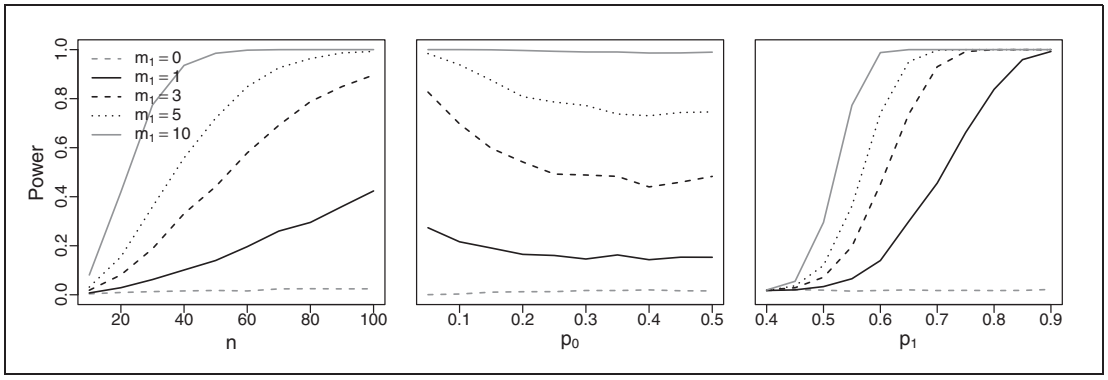

Figure 5. Power values for discrete data simulation for omnibus *log p* for $m_1 = \{0, 1, 3, 5, 10\}$, $m = 10$. The left chart shows results for increasing n and constant $p_0 = 0.4$, $p_1 = 0.6$; the chart in the middle, constant $n = 50$, increasing p_0 and $p_1 = p_0 + 0.2$; the right chart $n = 50$, $p_0 = 0.4$ and increasing p_1.

the power values are much lower than for equal mean effect sizes. Still, the omnibus *log p* method has maximum power in nearly all scenarios, only for $m_1 = 10$ the Fisher's combination test is more powerful.

Figure 4 (second row) shows results for negative effect sizes under the null hypothesis, leading to p-values that are stochastically larger than uniform. A comparison with Figure 2 reveals that this does not much influence the power of the omnibus test ($\Delta = 0.3\sqrt{m_1}$ for $m_1 \in \{1, 3, 10\}$, $m = 10$), but it reduces the power of the Fisher's combination test a lot for small m_1. The same is true for the Stouffer's test (not shown), as it also uses the sum over all (transformed) p-values. The power difference between the omnibus test and the Fisher's combination test reaches more than 70 percentage points, e.g. $m = 10$, $m_1 = 1$, $\Delta = 0.3$.

The power of the Bonferroni test and of HC changes even less compared to the omnibus test when parameters are in the interior of the null hypothesis.

If both alternative and null hypotheses have effect sizes distributed according to an exponential distribution (with a rate parameter of $3\sqrt{m_1}$ for alternative hypotheses and $-3\sqrt{m_1}$ for null hypotheses), the relative behavior of the methods (Figure 4, third row) is qualitatively similar to that implied in the second row. As observed in the first row, however, the power clearly decreases for all methods with randomly distributed effect sizes.

3.4 p-Values from discrete data

The assumption of uniformly distributed p-values under the null hypothesis is not always satisfied. Besides the possibility of parameter values in the interior of the null hypothesis, also discrete models lead to p-values that are not uniformly distributed on the interval [0, 1]. As p-values obtained from a discrete distribution are not covered by our underlying assumptions, we performed a simulation study to evaluate our test under such a situation. In view of our genetic application, we considered a two-sample binomial model. For large enough sample sizes, we would expect in general less effect of the discrete model. In our setup, however, type I error control was achieved even for small sample sizes.

For the first group, the simulated data were $B(n, p_0)$ distributed; for the second group, again generated from $B(n, p_0)$ under the null hypothesis and from $B(n, p_1)$ under the alternative. Here, n denotes the per-group sample size. A χ^2 test with one degree of freedom was performed and the corresponding p-value was calculated. If both groups showed only successes or only failures, the p-value was set to $p = 1$.

We first checked whether the type I error is still controlled under our discrete model. For this purpose, we considered sample sizes n between 10 and 100, as well as probabilities p_0 between 0.05 and 0.5 under the null hypothesis. Although the χ^2 test is usually not recommended for small expected cell frequencies, we nevertheless used the standard p-values produced by the R function *chisq.test*. Our simulations showed no violations of the type I error probability of $\alpha = 0.05$ with the omnibus test (see Figure 5). This may be since the χ^2 test tends to control the type I error probability fairly accurately even for small n.[16]

Figure 5 provides the power obtained when using our omnibus test on several scenarios for $m = 10$ and $m_1 \in \{1, 3, 5, 10\}$. The left plot shows the power values as a function of n from 10 to 100 for omnibus *log p* for $p_0 = 0.4$, and $p_1 = 0.6$. The plot in the middle shows the power values as a function of p_0 with constant $n = 50$ and

Table 2. Meta-analysis (example I).

	Omnibus				
	log p	Bonferroni	Simes	Fisher	HC
6 months	0.119	0.118	0.118	0.509	0.500
12 months	0.257	0.406	0.235	0.178	0.355
18 months	0.116	0.279	0.279	0.094	0.152
24 months	0.013	0.006	0.006	0.019	0.033

Note: Global tests have been applied to a meta-analysis comparing post-operative radiation therapy with or without adjuvant chemotherapy in patients with malignant gliomas. p-Values of the methods are shown when testing the global null hypothesis at different time points.

$p_1 = p_0 + 0.2$ increasing in the same amount as p_0. For the right plot, $p_0 = 0.4$, $n = 50$, and p_1 is increasing from 0.4 to 0.9.

4 Examples

4.1 Meta-analysis

In meta-analysis, the evidence from several studies on a topic is combined. There are several examples in the literature showing that the efficacy of a treatment can vary among studies. Reasons for such a variation can be, among other factors, due to the differences in the underlying study populations or environmental factors. If effect size estimates are available for all considered studies, a random effect meta-analysis is often carried out. Global tests, such as the Fisher's and the Stouffer's tests, are a popular alternative option that do not require effect size estimates.

As an illustration, we applied our omnibus test to a data set from a meta-analysis provided by the R-package metafor.[17] We chose the data set dat.fine1993 where results from 17 studies are presented which compare post-operative radiation therapy with or without adjuvant chemotherapy in patients with malignant gliomas.[18] For each study, the data set specifies the number of patients in the experimental group (receiving radiotherapy plus adjuvant chemotherapy) as well as the number of patients in the control group (receiving radiotherapy alone). In addition, the number of survivors after 6, 12, 18, and 24 months follow-up within each group is given. One of the 17 studies recorded survival only at 12 and 24 months. For illustration purposes, we performed a separate meta-analysis for each time point and calculated χ^2 test (or Fisher's exact test, where appropriate) for each study. The resulting p-values were then applied to test the global null hypothesis using the following methods: Bonferroni, Simes, Fisher, Stouffer, HC, and omnibus *log p*.

Table 2 shows the resulting p-values for the global tests. Note that the table does not display the results for Stouffer's method which in all cases results in a p-value close to 1 and will not be discussed further. As in the simulation study, the omnibus method is among the top methods for all time points except for the 12-month data, where the p-value of the Fisher's combination test is approximately one-third smaller than the p-values of the omnibus method. For the 6-month data, however, the advantage of the omnibus method as well as Bonferroni and Simes methods (all p-values between 0.12 and 0.13) over the Fisher's test (p-value: 0.51) is considerable. The largest p-value across all scenarios turns out to be smallest for the omnibus test.

We next analyzed the data examples from the R-package metap.[19] We used five of the eight different data examples, ignoring three that involve only hypothetical data. For each of these data sets, a vector of p-values of lengths ranging from 9 to 34 is provided in the package. For instance, the data taken from the meta-analysis by Sutton et al.[20] involve 34 randomized clinical trials where cholesterol lowering interventions were compared between treatment and control groups. The actual treatments were mostly drugs and diets. For each study, a test was performed to analyze if the effect sizes (log odds ratio) are smaller than 0 (one-sided test) and p-values were calculated based on the normal distribution (Sutton et al.,[20] Table 14.3). For details on the other data sets, we refer to the original publications and for references see the documentation of the metap package. Note that for some studies p-values were derived from independent subgroup analyses.

Table 3 compares different tests of the global null in terms of their p-values. Three of the methods (Simes, Fisher, *log p*) lead to significant p-values at level $\alpha = 0.05$ for four of the five data sets. The omnibus *log p* method, however, is the only test that also provides four significant results at level $\alpha = 0.01$.

Table 3. Meta-analysis (example II).

	Omnibus					
	log p	Bonferroni	Simes	Fisher	Stouffer	HC
Sutton	0.24	0.13	0.13	0.79	1	0.57
Mourning	0.007	0.07	0.04	0.017	0.11	0.013
Naep	<0.001	<0.001	<0.001	<0.001	<0.001	0.056
Teach	0.0007	0.019	0.019	0.0014	0.0077	0.24
Validity	<0.001	<0.001	<0.0001	<0.001	<0.001	0.025

Note: p-Values are obtained from several global null hypothesis tests. The data have been taken from the examples provided with the R-package *metap*.

4.2 Experimental evolution

With the development of large-scale inexpensive sequencing technologies, experiments became popular that aim to elucidate biological adaptation at the molecular level of DNA and RNA. In such experiments, organisms are often exposed to stress factors for several generations, and their genetic adaptation is studied. With microorganisms, such stress factors can for instance result from antibiotics, with the adaptation being resistance. With higher organisms, examples of stress factors are temperature or toxic substances. While evolution in nature usually takes place only once under comparable circumstances, experimental evolution can be done with replicate populations. Among other things, replication permits to investigate the reproducibility of adaptation, a key topic in evolutionary genetics. The statistical challenge is to identify genomic positions (called loci) involved in adaptation. There is a large number of candidate loci, for which adaptation has to be distinguished from random temporal allele frequency changes due to genetic drift as well as sampling and sequencing noise.

Furthermore, recent research suggests that replicate populations often do not show a consistent behavior, with signals of adaptation showing up partially at different loci. Two biological explanations for this finding are that beneficial alleles may be lost due to drift, and that the same adaptation at a phenotypic level can often be achieved in multiple ways at the genomic level.

When testing for significant allele frequency changes, a test like our omnibus test is therefore desirable, as it enjoys good power also when signals of adaptation are not consistent across replicates. We illustrate the application of our omnibus *log p* test to data from an experiment on *Drosophila* described in Griffin et al.[21] This experiment involves five experimental populations that are initially analyzed separately. For a given population, χ^2 homogeneity test has been computed for each of more than 2.5×10^6 candidate single nucleotide polymorphism (SNPs). As the standard p-values of the χ^2 test would only account for the sequencing variation, but not genetic drift and sampling, we obtained p-values using simulations under a null model that includes all sources of variation. Since the signals of selections were frequently not consistent across replicates, we then combined the matching p-values across the five replicate populations using our omnibus *log p* test. The combination of the p-values leads to 347,004 significant SNPs out of a total of 2.568×10^6 considered genomic positions at level $\alpha = 0.05$. Out of these SNPs, 5431 remain significant, after applying a Bonferroni correction. Controlling the false discovery rate (FDR) using the Benjamini–Hochberg procedure leads to a set S of 92,868 significant genomic positions. Due to genetic linkage, most of them are expected to be correlated with an influential SNP but not to be directly influential. Focusing on the FDR corrected p-values, 2241 of the SNPs in S were not significant with the Fisher's combination test. With the Bonferroni test (minimum of the five p-values for an SNP), 28,276 SNPs within S were not found.

A graphical summary of the p-values obtained using our omnibus test can be found in the online supplemental material.

5 Discussion

In this manuscript we introduced new non-parametric omnibus tests for testing the global null hypothesis. They require independent p-values as input and assume them to be uniformly distributed (or stochastically larger than uniform) under the null hypothesis. Our proposed approach enjoys very good power properties, no matter in how many cases the alternative holds. In our comparison with alternative approaches, it is not always the best method, but we did not find scenarios, where the omnibus test performs considerably worse than the best alternative

method for a given setup. One could furthermore construct better specialized tests in situations where knowledge is available concerning likely deviations from the global null. The proposed omnibus test is useful when no such information is available.

For our test, we compute successive cumulative sums of the suitably transformed sorted individual p-values. The most unusual cumulative sum is then obtained by computing the p-value of each sum under the global null hypothesis. The smallest p-value is then used as test statistic.

We consider different transformations of the initial p-values p_i, in particular $1 - p_i$, $-\log(p_i)$, z_{1-p_i}, and $p_i^{-1/2}$. Our results show only small differences in power between the transformations. However, the *log p* transform seems to lead to a particularly good trade-off in power across many scenarios.

As expected the Simes test outperforms the Bonferroni procedure (or is equal) in the simulation study, though, for the considered scenarios the improvement in power is not remarkable.

All our simulations are based on one-sided tests, but the methods also work for the two-sided testing scenario (see Figure 6 in the online supplemental material). For two-sided tests, however, it is also possible to reject the global null hypothesis even when the individual hypotheses show clear effects in differing directions.

Acknowledgements

We are grateful to P Griffin et al. for providing us with their experimental data and to Martin Posch for giving us helpful suggestions.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research of Andreas Futschik was supported in part by the National Science Foundation under Grant No. NSF PHY-1125915. Thomas Taus received funding from OEAW (DOC fellowship).

Supplemental material

Supplemental material for this article is available online.

ORCID iD

Sonja Zehetmayer http://orcid.org/0000-0001-6863-7997

References

1. Goeman JJ, Van De Geer SA, De Kort F, et al. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004; **20**: 93–99.
2. Ingster Y and Lepski O. Multichannel nonparametric signal detection. *Math Methods Statist* 2003; **12**: 247–275.
3. Fisher RA. *Statistical methods for research workers*. London: Oliver and Boyd, 1932.
4. Stouffer SA, Suchman EA, DeVinney LC, et al. *The American soldier: adjustment during army life*. Vol. 1, Princeton, USA: Princeton University Press, 1949.
5. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**: 751–754.
6. Tukey J. *T13: N the higher criticism. Course notes. Statistics 411*. Princeton: Princeton University Press, 1976.
7. Donoho D and Jin J. Higher criticism for detecting sparse heterogeneous mixtures. *Ann Statist* 2004; **32**: 962–994.
8. Donoho D and Jin J. Higher criticism for large-scale inference, especially for rare and weak effects. *Statist Sci* 2015; **30**: 1–25.
9. Stigler SM. Linear functions of order statistics. *Ann Math Stat* 1969; **40**: 770–788.
10. Crocetta C and Loperfido N. The exact sampling of L-statistics. *Metron* 2005; **63**: 1–11.
11. Nagaraja HN. *Order statistics from independent exponential random variables and the sum of the top order statistics*. Boston, MA: Birkhäuser, 2006, pp.173–185.

12. Dickhaus T. *Simultaneous statistical inference. With applications in the life sciences*. Springer: Heidelberg, 2014.
13. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B* 1995; **57**: 289–300.
14. R Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, www.R-project.org/ (2017, accessed 19 March 2018).
15. Haibe-Kains B, Desmedt C, Sotiriou C, et al. A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics* 2008; **24**: 2200–2208.
16. Parshall CG and Kromrey JD. Tests of Independence in contingency tables with small samples: a comparison of statistical power. *Educ Psychol Meas* 1996; **56**: 26–44.
17. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw* 2010; **36**: 1–48.
18. Fine HA, Dear KB, Loeffler JS, et al. Meta-analysis of radiation therapy with and without adjuvant chemotherapy for malignant gliomas in adults. *Cancer* 1993; **71**: 2585–2597.
19. Dewey M. *metap: meta-analysis of significance values*. R package version 0.8, 2017.
20. Sutton AJ, Abrams KR, Jones DR, et al. *Methods for meta-analysis in medical research*. London: John Wiley and Sons, Ltd, 2000.
21. Griffin PC, Hangartner SB, Fournier-Level A, et al. Genomic trajectories to desiccation resistance: convergence and divergence among replicate selected drosophila lines. *Genetics* 2017; **205**: 871–890.