Bioinformatics, 32(6), 2016, 850–858
doi: 10.1093/bioinformatics/btv608
Advance Access Publication Date: 16 November 2015
Original Paper

OXFORD

Genetic and population analysis

Prioritizing hypothesis tests for high throughput data

Sangjin Kim and Paul Schliekelman*

Department of Statistics, University of Georgia, Athens, GA 30602, USA

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on December 22, 2014; revised on September 11, 2015; accepted on October 16, 2015

Abstract

Motivation: The advent of high throughput data has led to a massive increase in the number of hypothesis tests conducted in many types of biological studies and a concomitant increase in stringency of significance thresholds. Filtering methods, which use independent information to eliminate less promising tests and thus reduce multiple testing, have been widely and successfully applied. However, key questions remain about how to best apply them: When is filtering beneficial and when is it detrimental? How good does the independent information need to be in order for filtering to be effective? How should one choose the filter cutoff that separates tests that pass the filter from those that don't?

Result: We quantify the effect of the quality of the filter information, the filter cutoff and other factors on the effectiveness of the filter and show a number of results: If the filter has a high probability (e.g. 70%) of ranking true positive features highly (e.g. top 10%), then filtering can lead to dramatic increase (e.g. 10-fold) in discovery probability when there is high redundancy in information between hypothesis tests. Filtering is less effective when there is low redundancy between hypothesis tests and its benefit decreases rapidly as the quality of the filter information decreases. Furthermore, the outcome is highly dependent on the choice of filter cutoff. Choosing the cutoff without reference to the data will often lead to a large loss in discovery probability. However, naïve optimization of the cutoff using the data will lead to inflated type I error. We introduce a data-based method for choosing the cutoff that maintains control of the family-wise error rate via a correction factor to the significance threshold. Application of this approach offers as much as a several-fold advantage in discovery probability relative to no filtering, while maintaining type I error control. We also introduce a closely related method of *P*-value weighting that further improves performance.

Availability and implementation: R code for calculating the correction factor is available at http://www.stat.uga.edu/people/faculty/paul-schliekelman.

Contact: pdschlie@stat.uga.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A dominant trend in biology in recent years has been the development of high throughput techniques and the dramatic increase in the resolution of available data. However, most of the information gained is not relevant for any particular question at hand and comes at the cost of more hypothesis tests and thus more stringent statistical thresholds. There is often high redundancy between tests and the gain in information may be slower than the increase in resolution. Thus, higher resolution will not always lead to higher probability of discovery.

Given the realities of multiple testing, it is unlikely that a mere increase in throughput and resolution will greatly increase discoveries. Rather, it will be necessary to combine high throughput data with other sources of information in order to better target investigations. The problems of multiple testing are well understood and many methods have been proposed for using external information

to filter for the most promising features of the data. Such methods typically have two stages: first, some filtering criterion is used to select the most promising features. Then, only those features are tested for the effect of interest. These include methods for microarrays (Bourgon, *et al.*, 2010; Clevert, *et al.*, 2013; Hackstadt and Hess, 2009; Jiang and Doerge, 2006; Lu, *et al.*, 2011; McClintick and Edenberg, 2006; Talloen, *et al.*, 2007; Talloen, *et al.*, 2010), RNA-Seq (Bottomly, *et al.*, 2011; Ramskold, *et al.*, 2009; Rau, *et al.*, 2013a, b; Sultan, *et al.*, 2008), genome-wide association studies (Calle, *et al.*, 2008; Dai, *et al.*, 2012; Degnan, *et al.*, 2008; Li, *et al.*, 2013; Patwardhan, *et al.*, 2014; Roeder and Wasserman, 2009; Van Steen, *et al.*, 2005) and epistasis (Emily, *et al.*, 2009; Evans, *et al.*, 2006; Pattin and Moore, 2008; Yang, *et al.*, 2009).

Despite the popularity of filtering methods, key questions remain about their general statistical properties. Bourgon *et al.* (2010) discussed the conditions sufficient for maintaining Type I error control and showed that the key requirement is that the null hypothesis distribution of the test statistics after filtering should be the same as the null hypothesis distribution before filtering. They showed that some filtering techniques in use can violate this requirement. Their focus was on the conditions for filtering to be valid. Little is known about the conditions required for filtering to be successful in significantly increasing discovery probabilities, and our focus is on this question.

In this paper, we address two major issues. First, we evaluate the usefulness of filtering and determine major factors affecting its behavior. We quantify the effect of the filtering statistic in terms of its probability of ranking true effects highly. We show that a strong filter that has a high probability of ranking true positives in e.g. the top 10% can greatly increase discovery probability (as much as 20-fold) when there is high redundancy between features and when a good cutoff is known in advance. Even a random filter can increase power when the cutoff point is well chosen. On the other hand, filtering is less effective when there is low redundancy between features and if the filter statistic is not able to reliably rank true positive features highly.

Second, most applications of filtering methods have used ad-hoc approaches to choosing the filter cutoff point. The filter cutoff point refers to the value of the filtering criterion which separates features that will be included in the second stage and those which will not. We show that the filter cutoff has a large effect on the performance of filtering methods. A good choice can make a several-fold difference in discovery probability relative to a poor choice. Furthermore, inappropriate ad-hoc methods can greatly inflate false positive rates. We introduce a general and rigorous method for choosing the filter cutoff and show that this approach can increase discovery probabilities by several-fold relative to no filtering. We also introduce a simple and intuitive method for weighting p-values that is closely related to our filtering method and improves the performance further.

2 Methods

2.1 Optimized filtering

Consider a generic high throughput study. There are L true effects in the data and m potential hypothesis tests that could be performed in order to find those effects. In a QTL mapping study, the true effects would be genetic polymorphisms that affect the trait of interest and the hypothesis tests would be genetic markers. In an RNA-seq study, the true effects would be genes that are differentially expressed between treatment conditions and the hypothesis tests would be the genes that are tested.

Now, suppose that the hypothesis tests are ordered by a filtering criterion that attempts to identify the tests most likely to be able to identify true effects with the best candidate test being ranked #1, second best #2 and so on. The filtering criterion will be based on some biological knowledge of the hypothesis tests derived from independent data. The top v tests on this filter-ordered list will be conducted and the remaining tests will be discarded. Discarding tests will reduce multiple testing and potentially increase power. Our goal is determining what v should be in order to maximize power.

We choose a vector $\vec{n} = (n_1, \ldots n_R)$ of proposed values for v, where n_i is the number of filter-ordered tests, R is the total number of subsets of tests, $n_1 < n_2 <, \ldots, < n_R \leq m$ and each subset is included in the larger ones. Thus, we test 1 to n_1, 1 to n_2, etc. Take $r_1, r_2, \ldots, r_m$ to be the P-values for the filter-ordered hypothesis tests. The test set n_k refers to the filter-ordered set of hypothesis tests 1 to k. A hypothesis test i in test set n_k is significant if

$$r_i < \frac{\alpha}{\lambda \times n_k} \tag{1}$$

where α is the desired family-wise error rate (FWER). That is, a Bonferroni correction for the number of tests n_k is applied with a correction factor λ that will correct the multiple testing adjustment for the fact that we will optimize v over $\vec{n}$. We then take whichever test set maximizes the number of rejected null hypotheses. We show in the SI that

$$\text{FWER} \leq 1 - \left(1 - \frac{\alpha}{\lambda \times n_1}\right)^{n_1} \prod_{i=1}^{R-1} \left(1 - \frac{\alpha}{\lambda \times n_{i+1}}\right)^{n_{i+1} - n_i} \tag{2}$$

This has a simple interpretation: in order for no tests to be significant, then no p-value can be lower than the least-stringent threshold at which it is tested. The least-stringent threshold to which the first n_1 tests are compared is $\alpha/(\lambda \times n_1)$. The least stringent threshold to which the next n_2–n_1 tests are compared is $\alpha/(\lambda \times n_2)$,etc. We obtain the correction factor λ required to control FWER by equating (2) to α and solving for λ.The resulting equation does not have a closed form solution, but can easily be solved numerically. We conducted simulations to verify that this expression correctly controls FWER (Supplementary Fig. S1). Supplementary Figures S1 and S2 show a plot of calculated correction factors as a function of the number of elements of $\vec{n}$. The Supplementary Material also shows approximate closed-form solutions for the correction factor.

We focus on the control of FWER in this paper, which is important for applications such as QTL and genetic association mapping. A future paper will address control of FDR.

Some filtering schemes will result in tied ranks. Provided that tied tests lie in the same block between consecutive elements of the vector $\vec{n}$, then they will be tested at the same significance threshold.

Our approach assumes independence between tests in calculating the correction factor. However, there will usually be substantial correlation between tests. An alternative is to estimate an effective number of independent tests. Such approaches can be implemented in our method by calculating an effective number of tests within each subset. Simulation results in the SI show that this approach can improve power while maintaining Type I error.

We illustrate the above method with a simple example in Table 1. There are $m = 9$ total hypothesis tests and we take $\vec{n} = [3, 6, 9]$. The filter-ranked P-values are shown in the table. Note that the filter-ranking will not correspond exactly to the p-value order and thus the p-values are not ordered. First consider the case with no correction to the significance threshold. In the first test set, the P-values are tested against a threshold of $0.05/3 = 0.0167$ and two

Table 1. Toy example with $m = 9$ *P*-values and test sets of $\vec{n} = [3, 6, 9]$

P-value	Without correction factor ($\lambda = 1$)			With correction factor ($\lambda = 1.793$)		
	Test set			Test set		
	$n_1 = 3$	$n_2 = 6$	$n_3 = 9$	$n_1 = 3$	$n_2 = 6$	$n_3 = 9$
Significance Threshold	.0167	.0083	.0056	.0093	.0046	.0031
P_1	**.0011**	**.0011**	**.0011**	**.0011**	**.0011**	**.0011**
P_2	**.0092**	.0092	.0092	**.0092**	.0092	.0092
P_3	.0201	.0201	.0201	.0201	.0201	.0201
P_4		.0089	.0089		.0089	.0089
P_5		.0091	.0091		.0091	.0091
P_6		**.0064**	.0064		.0064	.0064
P_7			**.0022**			**.0022**
P_8			.0861			.0861
P_9			**.0045**			.0045

The significance thresholds are calculated as $0.05/n_i$ without the correction factor and $0.05/(n_i \times 1.793)$ with the correction factor. The *P*-values in bold are less than the significance threshold for that test set.

are significant (shown in bold). In the second test set, the threshold becomes $0.05/6 = 0.0083$. One of the previously significant *P*-values 0.0092 becomes non-significant, but a newly added *P*-value 0.0064 is significant. In the third test set, the *P*-values are tested against a threshold of $0.05/9 = 0.0056$. Now the 0.0064 *P*-value drops out of significance, but two new *P*-values 0.002 and 0.0045 are significant. The third test set has the highest number (three) significant *P*-values and thus we would take $\nu = 9$ as the optimal cutoff.

However, we have not accounted for the effect of this optimization on the FWER. We are maximizing rejections over three sets and the expected number of false positives will be increased by this maximization. The probability of type I error is the probability that there is at least one false positive across the three sets. The first set in which a *P*-value appears has the least stringent threshold that p-values will be tested against. If it is not rejected in that test set, then it will not be rejected in any (see SI for more details). Thus, the probability of type I error is

$$\begin{aligned} 0.05 &\geq P(S \geq 1) \\ &= 1 - P(S = 0) \\ &= 1 - (1 - 0.05/3\lambda)^3 (1 - 0.05/6\lambda)^{(6-3)} (1 - 0.05/9\lambda)^{(9-6)} \end{aligned}$$

where 0.05 is a target FWER, S is the number of type I errors and λ is the correction factor based on 3 candidate subsets. We numerically solve this equation for λ, yielding $\lambda = 1.793$. After applying the correction factor, the significance thresholds decrease to the values shown in the table and the number of significant tests decreases as well. Now, test sets 1 and 3 both have two significant tests and thus we can take either $\nu = 3$ or $\nu = 9$ as the optimal cutoff.

2.2 Power gains from filtering

Our first goal is to explore the potential gain in statistical power from filtering. Consider again the generic high-throughput experiment discussed earlier. We will focus on one specific true effect that we label as effect τ. There are m potential hypothesis tests that could be performed in order to find this effect and r of these tests can actually detect the effect τ. 'τ-tests' refers to these tests. We will make the approximation that all of these r tests have the same power to detect τ.

Suppose that all of the tests are ordered by a filtering criterion and the first ν such tests are conducted and the remainder discarded. For the moment, we will assume that ν is fixed in advance so that a simple Bonferroni correction will suffice and no adjustment to the significance criterion needs to be made. The probability that at least one of the r τ-tests successfully detects τ is

$$\psi(\nu) = \sum_{s=1}^{r} \gamma_\nu(s)\varphi_\nu(s)$$

where ν is the number of filter-ordered hypothesis tests conducted, $\gamma_\nu(s)$ = probability that s out of the r τ-tests are included in the top ν tests, and $\varphi_\nu(s)$ = probability that at least one of the s τ-tests detects τ. We refer to $\psi(\nu)$as the discovery probability. We use this terminology to distinguish it from the power for a single test. We assume that the test statistics for the r τ-tests follow a multi-normal distribution with mean 0 when the null hypothesis is true and with the mean determined by the effect size when the null hypothesis is false. The correlation coefficients will be varied and reflect the amount of correlation between tests.

The function $\gamma_\nu(s)$quantifies the effectiveness of the filter. An effective filter will rank hypothesis tests highly that are able to detect true effects. We assume that the top ν filter-ordered tests are independently sampled without replacement from the m total tests, of which r are τ-tests. Take $q(u)$as the probability that a given τ-test is included in the top proportion u of all hypothesis tests. $q(u)$ is modeled with a beta CDF, giving great flexibility for determining the properties and effectiveness of the filter function. If sampling was with replacement, then the function $\gamma_\nu(s)$ would follow the binomial distribution with ν trials and $q(\nu/m)$ as the probability of success. However it instead follows Wallenius' noncentral hypergeometric distribution (Fog, 2008a, b) because sampling is without replacement. If the filtering is effective, then the τ-tests will have a higher probability of being selected, which distinguishes the distribution from a standard hypergeometric. Wallenius' noncentral hypergeometric accounts for this. See SI for further details.

The assumption of independence in filter rank may be violated in some data sets and the benefits of filtering will be lower (see SI). The discovery probability formula is implemented in an R program that is included in the SI.

Example Scenarios. In our calculations below, we will consider two basic scenarios. The first scenario has a single test capable of detecting each true effect. This is characteristic of, for example, microarray and RNA-seq studies. In these studies, we are interested in determining which genes are differentially expressed between

biological conditions. Typically, each gene is its own hypothesis test and true effects are genes that are differentially expressed. We assume 10 000 hypothesis tests (that is, 10 000 genes being tested). This is an example with no redundancy between tests. That is, each true effect (differentially expressed gene) can be detected only by a single test.

In the second scenario there are 20 tests capable of detecting the true effect and the tests are correlated. This is characteristic of, for example, QTL mapping, where each marker is a hypothesis test and the true effect is the QTL. Typically, there are multiple markers that are correlated with the QTL and each other. We assume 1000 hypothesis tests (that is, 1000 markers being tested). In this case, there is redundancy between tests because multiple markers can detect the same true effect. The level of redundancy depends on the amount of correlation between tests.

2.3 Weighted *P*-values

Next, we consider a closely related approach based on the weighted *P*-value framework (Benjamini and Hochberg, 1997; Finos and Salmaso, 2007; Genovese, *et al.*, 2006; Holm, 1979; Kropf, *et al.*, 2004; Roeder and Wasserman, 2009; Roquain and van de Wiel, 2009; Rubin, *et al.*, 2006; Wasserman, *et al.*, 2006; Westfall, *et al.*, 2004). Suppose that we have *P*-values p_1 to p_m (not ordered). A weight w_j is calculated for each *P*-value p_j such that the corresponding null hypothesis is rejected if

$$\frac{p_j}{w_j} < \frac{\alpha}{m}$$

Equation (1) above suggests the following weight for the jth filter-ordered *P*-value $r_{j:}$

$$w_j = \frac{m}{\lambda \times n_{k(j)}} \tag{3}$$

where $n_{k(j)}$is the smallest value from the vector $\vec{n}$ that is greater than j and λ is the correction factor computed from Eq. (2). This leads to a different procedure than the optimal filter approach proposed above. In the optimal filtered approach, we maximize over the values of $\vec{n}$. That is, a test is only significant if its filter rank is less than n_{opt} and its *P*-value is less than $\alpha/\lambda \times n_{\text{opt}}$, where n_{opt} is the value of $\vec{n}$ for which the number of significant tests is maximized. In contrast, the optimization step does not occur in the weighted *P*-value approach and a test is rejected if its *P*-value is less than $\alpha/\lambda \times n_{k(j)}$. Thus, there are potentially more rejected null hypotheses in the second procedure. We call this block weighting, where blocks refer to the groups 1 to n_1, n_1+1 to n_2, etc.

Consider the example in Table 1. After applying the correction factor, we found two significant tests in both the first and third test sets. Under the optimal filtering procedure, we would choose one of these test sets as optimal and have two significant tests. Under the block weighting procedure, the first three *P*-values are compared to 0.0093, the second three are compared to 0.0046 and the third three are compared to 0.0031. All tests that meet their respective significance criterion are rejected and thus there are three significant tests.

Surprisingly, these two procedures have the same FWER bound. Although the weighted *P*-value approach will clearly produce more false positives under some scenarios, the probability of producing zero false positives is the same between the two procedures and FWER is based on this probability (see SI). Because there is no downside from the FWER perspective in using the block weighting scheme and it will produce more true positives in some scenarios, then it is advantageous to use the weighted *P*-value scheme (3) instead of the optimal filtering one. We will return to this point in the data application. This weighting scheme can be viewed as a generalization of that of Ionita-Laza *et al.* (2007), who uses $n_1 = \text{k}$, $n_i = n_{i-1} + 2^{i-1}k, i = 2, \ldots$ for some integer k and then weight appropriately.

3 Results

3.1 Random filter

We start with a random filter. That is, v of the total m tests are randomly chosen, with all tests being chosen with equal probability regardless of whether they are null or non-null. Figure 1 shows the discovery probability as a function of v for several scenarios. The first is parameterized ($r = 1$, $m = 10\,000$) for the RNA-seq type experiment described above. We see that the discovery probability increases rapidly with v initially, but then flattens out as v increases past about 1000 and then increases slowly after that. A 5-fold increase of v from 2000 to 10 000 results in a roughly 2-fold increase in discovery probability.

There are two opposing forces in action as v is increased. The probability that the τ-test is included in the top v is v/m. If this was the only effect in play, then doubling the number of tests would double the probability of discovery. However, the statistical power for each test decreases as the number of tests increases. Thus, the shape of the curve in Figure 1a is the result of the interaction of these two effects. The initial rapid increase is due to the linear increase in v/m. However, the decreasing power per test causes the curve to flatten out.

Figure 1b–d are parameterized for the QTL mapping scenario described previously ($r = 20$, $m = 1000$). Figure 1b has low correlation (0.2) between the 20 τ-tests. The dynamics are similar to those in Figure 1a. However, the beneficial component of increasing v is saturating. That is, the higher v is, then the higher the probability is

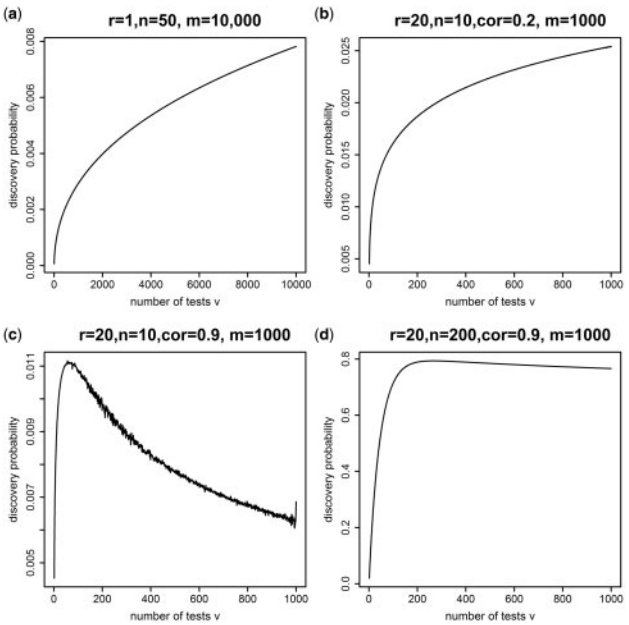

Fig. 1. Plots of the probability of detecting a true effect with a random filter. *r* is the number of hypothesis tests able to detect the effect, *cor* is the pairwise correlation between those tests, and *n* is the sample size. The probability of a test being in the top *v* is *v/m*. Waviness in the curves is due to numerical error in the R routine *pmvnorm* for calculating multivariate normal probabilities

that one of the 20 τ-tests has already detected the effect and therefore the benefit in increasing ν is lowered. Thus, after the initial period of rapid increase, the discovery probability curve increases even more slowly than in Figure 1a. A 5-fold increase in the number of tests from 200 to 1000 hypothesis tests results in only a 40% increase in discovery probability.

The impact of this saturation is more dramatic with high correlation between tests. Figures 1c has identical parameters to Figure 1b, except that the correlation between τ-tests is 0.9. Such a level of correlation is consistent with high density markers in QTL mapping, for example in genotype-by-sequencing studies. In this case, the discovery probability reaches a peak at 78 tests and thereafter decreases with increasing ν to about half of peak discovery probability. When the correlation between τ-tests is high, then there is a high level of redundancy between tests. Increasing ν beyond this point is counter-productive because the information gain is minimal and is outweighed by the increased multiple testing correction.

Figure 1d is identical to Figure 1c, except that the sample size is increased to 200. We see that discovery probability reaches a peak of about 80% at $\nu = 200$ and then decreases. However, in this case the decrease is minimal and the discovery probability does not drop below 75%. When the power for single tests is high, then the effect of the multiple testing correction is much lower than when the power is low. If the test statistic follows a normal distribution, then the power for a test is $1 - \Phi\left(\Phi^{-1}(1-\alpha/\nu) - \xi\right)$, where Φ is the standard normal cdf, ξ is the standardized effect size and $\Phi^{-1}(1-\alpha/\nu)$ is the Bonferroni-corrected significance threshold. When ξ is small, then changes in significance threshold from increasing ν are a dominating effect. However, when ξ is large, then changes in the significance threshold cause only a small change in $\Phi^{-1}(1-\alpha/\nu) - \xi$. It is primarily the low power situation that is of interest in this paper. When power is high, then special efforts to optimize discovery probability are not needed. Figures 1a–c have peak discovery probabilities on the order of 1%. In many situations of practical interest, discovery probabilities are of this order or lower. The interaction between effect size and filtering effectiveness is explored more thoroughly in the SI.

In summary, we see that a random filter can be effective for applications such as QTL mapping where there is high redundancy between tests, but will be detrimental for applications such a transcriptomics where there is minimal redundancy between tests.

3.2 Non-random filter

Next, we consider the case of filters that are not random, but rather are able to preferentially identify tests that are associated with true effects. We quantify the effectiveness of the filter by the probability $q(u)$ that a given τ-test is included in the top proportion u of all hypothesis tests.

Figure 2a shows the probability density functions for three different filter functions that we will consider. The corresponding cdfs give the probability $q(u)$ for a test to be ranked in the upper proportion u of tests. These are each modeled with a beta distribution with parameters as shown in the figure legend. The $\alpha = 1,\ \beta = 25$ is a very effective filter that has a maximum probability density at $u = 0$ and 93% probability of ranking a τ-test in the top 10%. The $\alpha = 1, \beta = 10$ curve is a less strong but still effective filter that has a 65% probability of ranking a τ-test in the top 10% of tests. The third filter $\alpha = 2, \beta = 10$ has peak probability at 0.1. It has a low probability of ranking a τ-test very highly, but has a high probability (67%) of ranking them in the top 20%. These filter-functions span a range from highly effective at increasing discovery probabilities to only marginally effective or even detrimental (see Figs 2 and 4).

The other three panels show the discovery probabilities for these filters under the same three low power scenarios considered in Figure 1. A strong filter can increase discovery probability greatly compared to no filtering (which occurs when $\nu = m$ at the extreme right end of the plot). With the optimal cutoff point, the $\alpha = 1, \beta = 25$ filter increases discovery probability by a factor of 19 for the high correlation case, a factor of 7 for the low correlation case, and a factor of 4 for the $r = 1$ case. The $\alpha = 1, \beta = 10$ increases power by a factor of 9 for the high correlation case and a factor of 4 for the lower correlation case, and a factor of 2.5 for the $r = 1$ case. The third filter ($\alpha = 2, \beta = 10$) increases discovery probability by 2- to 4-fold over the three scenarios.

Unlike with the random filter case, these filters are effective in increasing discovery probabilities for both the low-redundancy transcriptomic scenario and the high redundancy QTL mapping scenario. However, the benefit is much greater for the QTL mapping scenario.

The performance of the filters is strongly dependent on the choice of ν. The filter has a sharp peak, especially in the high correlation case. For this case, the peak is at $\nu = 3$. Discovery probability is 40% of the maximum at $\nu = 50$ and 25% of maximum at $\nu = 100$. The peaks are less sharp for other cases, but in most cases the discovery probability is heavily influenced by the choice of ν. The sharp peaks in the high correlation case (Fig. 2d) are at least partially caused by the assumption of independence in filter-ranks between tests. Results presented in the SI explore the effect of this assumption.

3.3 Optimizing the filter

The above results show that a properly chosen filter can greatly increase the probability for detecting target effects. However, these results assume that we know the optimal filter cutoff in advance.

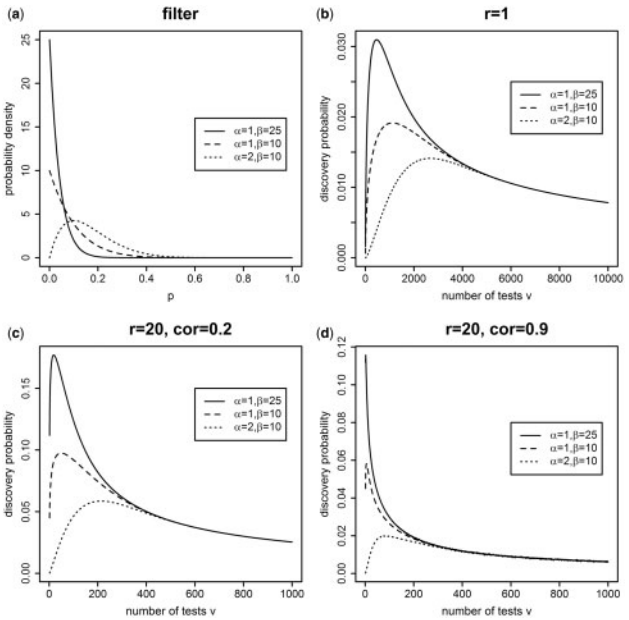

Fig. 2. Effectiveness of Filter. (**a**) The three different filter functions. The *x*-axis quantity *P* is the quantile at which a true effect test is ranked. The *y*-axis is the probability density for that ranking. (**b–d**) The discovery probabilities for these three filters for three different scenarios. The *x*-axis is the number of ranked tests conducted and the *y*-axis is the discovery probability

Unfortunately, the optimizing process will inflate Type I error unless properly accounted for. We demonstrate this via simulation in Figure 3. In this simulation we assume that we search for the optimal filter cutoff v by trying different values and choosing the one that gives the highest number of significant tests. We see that the Type I error increases dramatically as we increase the number of values of v that we optimize across (see SI for more details).

In the Section 2, we introduced a method for choosing the best value of v in a statistically principled fashion. We will examine its effect on discovery probabilities here. A key factor in applying this method is the set of values over which v is optimized. Choosing the best set of values is not straightforward. The sharper the peak in the discovery probability with respect to v, the more benefit in searching over a finer grid. However, this comes at the cost of a higher correction factor for type I error. We will not delve deeply into this issue here, but show one example in Figure 4. This is identical to Figure 2, but with a correction factor of $\lambda = 2.86$. This corresponds to searching v from 500 to 5000 in increments of 500 for the case with 10 000 total tests (Fig. 4b) or searching v from 50 to 500 in increments of 50 for the case with 1000 total tests (Fig. 4c, d). The vertical lines in each plot show the closest value of v to the peak value that will be obtained in this search scheme. The horizontal line in each plot shows the discovery probability with no filtering.

Note that while the shape of these curves is identical to Figure 2, the height of the curves has been reduced by the introduction of the correction factor. The weaker filter $(\alpha = 2, \beta = 10)$ has little or no benefit after the optimization over v is accounted for. However, the other two filters still bring major benefit. For the $r = 1$ scenario, the best searched value of v comes very close to the true optimum for all three cases. For the best $(\alpha = 1, \beta = 25)$ filter, there is an increase in discovery probability of 2.2-fold. For the second best filter $(\alpha = 1, \beta = 10)$, there is an increased in discovery probability of 35%. The optimized filters are more effective for the $r = 20$ cases. With correlation between tests of 0.2, the best searched value of v hits exactly the true optimum for the $(\alpha = 1, \beta = 10)$ filter, but misses substantially for the $(\alpha = 1, \beta = 25)$ case (true optimum at $v = 19$, searched peak at $v = 50$). The increase in discovery probability relative to no filter is 3.08-fold for the $(\alpha = 1, \beta = 25)$ and 1.82-fold for the $(\alpha = 1, \beta = 10)$ case. The peaks are very sharp for the high correlation case. The best searched values of $v = 50$ miss the true optimums of 2 and 5 for the $(\alpha = 1, \beta = 25)$ and $(\alpha = 1, \beta = 10)$ filters by a large margin. Still, the increase in discovery probability relative to no filtering is 3.4-fold and 2.7-fold, respectively. It is clear that we could do better if we knew something about the shape of the filter function and therefore how finely to search v. In absence of this information, the decision is more difficult and the filter effectiveness is likely to be decreased.

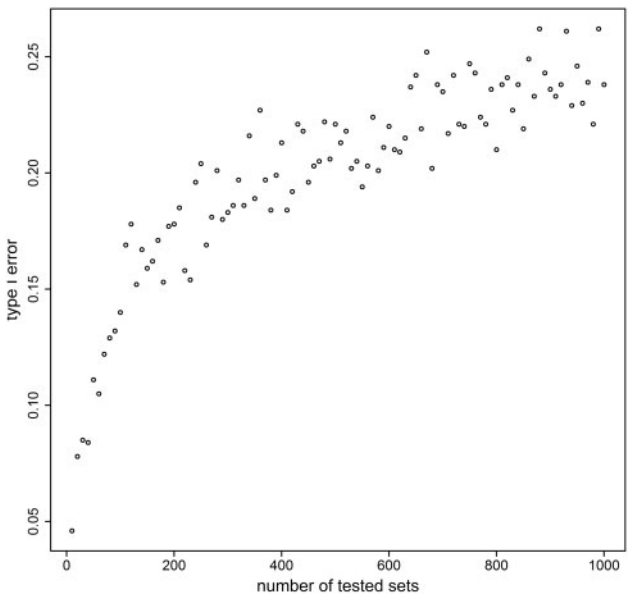

Fig. 3. Type I error rate at the level of $\alpha = 0.05$ in naïve optimization of the filter cutoff. The *y*-axis is type I error. The *x*-axis is the number of tested sets. For example, if there were 1000 total tests and we tried v=100, 200, . . . , 1000, then there would be 10 sets tested. Each point represents the proportion of rejected null hypotheses in 1000 simulation replicates, with *P*-values generated from a uniform distribution

3.4 Application of the filtering/correction factor procedure to a mouse obesity data

We applied our filtering approach to the QTL mapping data set of Ghazalpour *et al.* (2006), using gene expression information to rank the markers. The data set includes 1065 SNP markers, expression values for 3421 genes, and measurements of several phenotypes including body weight for 135 female mice. The goal is to identify QTLs for body weight. We used the Mouse Gene Expression Database(Smith, *et al.*, 2014) to identify a set of 37 genes present in the data that have previously been shown to be associated with body weight in mice. LOD scores were calculated for each SNP with respect to each of these 37 genes. The filter statistic for each SNP was the median value over the 37 genes. These were then sorted to give filter ranks. The SNPs were then tested for linkage with body weight QTLs. See the SI for more details.

The reasoning in using this filtering function is that SNPs that drive expression of genes associated with body weight are good

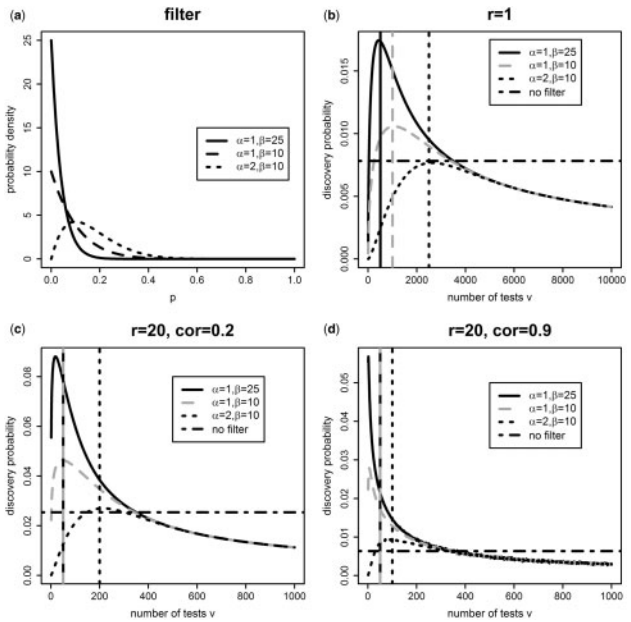

Fig. 4. Effectiveness of filter after optimizing. (**a**) The three different filter functions. (**b–d**) The discovery probabilities for these three filters for three different scenarios, with a correction factor of $\lambda = 2.86$. This correction factor corresponds to optimizing v over the values 500, 1000, 1500, . . . , 5000 for (b) and 50, 100, 150, . . . , 500 for (c) and (d). The horizontal line shows the discovery probability with no filtering. The vertical line shows the closest value of v to the peak that will be obtained under the search scheme, with the line types of the vertical lines being the same as the filter to which they correspond

candidates for being body weight QTLs. Because we did not use the body weight data in calculating the filter ranks, then the filter ranks will be independent of the LOD scores with respect to body weight under the null hypothesis that the SNP is not a body weight QTL. We consider both the optimal filtering approach (Eq. (1)) and the *P*-value weighting approach (Eq. (3)).

There were four chromosomal regions with some evidence for body weight QTLs, although no SNP is significant under the standard Bonferroni correction with 1065 tests. For the SNP with the best chance of being detected in each of the four QTL regions, Table 2 shows the filter rank and the maximum number of tests for which it would be significant under a Bonferroni correction. The maximum number of tests gives a measure of how significant the test is. If the maximum tests equals or exceeds $m = 1065$, then the SNP would be significant under a standard Bonferroni correction. The lower this quantity is below m, then the better the filter will have to perform in order for the SNP to be significant.

The SNP shown in the table for each QTL region is the SNP with the highest value of (max tests to be significant)/(filter rank). The most significant SNPs are on chromosome 19. The strongest would be significant if tested among 591 or fewer SNPS. These SNPS have filter ranks of about 400 (among 1065). Three other SNPs in the same region are moderately weaker, but are filter-ranked at about the 17th percentile. The most significant SNP of these is shown in the Table 2. It would be significant if tested among 444 or fewer markers. The next most significant region is on chromosome 15. The most significant markers would be significant if tested among 190 or fewer markers. These markers are filter-ranked very highly at about the 1% point. The third most significant region is on chromosome 5. The strongest SNP would be significant if tested among 124 or fewer markers. The filtering ranking is ineffective with this SNP, placing it about the 33rd percentile. The fourth strongest region is on chromosome 1. Although several markers are filter-ranked at about the 10th percentile, the most significant SNPs would have to be filter-ranked in the top 5% to be significant. No other regions show any evidence of QTLs. The most significant SNP outside of these four regions would only be significant if tested among six or fewer markers and most of the remaining *P*-values are greater than 0.05. Supplementary Table S1 in the SI shows the top 20 most significant SNPs.

Table 2 shows the results of using three different schemes for optimizing the filter: $\vec{n} = 100$–500 in increments of 100, $\vec{n} = 25$–1050 in increments of 25 and $\vec{n} = 50$–200 in increments of 50. For each combination of SNP and search scheme, we show the lowest value (labeled 'Next highest inc.') of $\vec{n}$ that is higher than the filter rank (i.e. the value at which the SNP will be tested), and that value multiplied by the correction factor (i.e. the effective number of hypothesis tests in the Bonferroni correction for that SNP, labeled 'Corrected'). The SNP will be significant if this corrected value is less than or equal to the maximum Bonferroni-corrected number of tests for that SNP (column 3 in the Table). The corrected values are bold in cases where this occurs.

Under the first scheme, the QTL on chromosome 19 is significant. It has filter rank 182 and thus the next highest increment is 200. After multiplying by the correction factor of $\lambda = 2.22$, the effective number of tests is 444 and the marker is significant. No other markers are significant. Although the marker on chromosome 15 is very highly ranked (12th), the next highest increment is 100 under this relatively coarse search scheme. This marker is too weak to be significant under the resulting 222 effective number of tests. These results are the same whether we use the optimal filtering approach or the weighted P-value approach.

The second search scheme uses increments of 25. This causes the correction factor to increase from 2.22 to 4.17. However, this results in a decrease from 100 to 25 in the next highest increment for the 12th-ranked chromosome 15 marker. Even after applying the higher correction factor, the marker is still significant. On the other hand, the chromosome 19 marker is no longer significant. The finer search grid makes no difference in the next highest increment (200 in both cases) and the higher correction factor causes the marker to lose significance.

The third search scheme uses increments of 50, but only for the top 200 filter-ordered tests. The 12th-ranked chromosome 15 marker has a next highest increment of 50. After applying the correction factor of 2.03 the effective number of tests is 104 and it is significant. The 180th-ranked chromosome 19 marker again has a next highest increment of 200. With the lower correction factor of 2.03, the effective number of tests is 406 and it is also significant. Under the optimal filtering approach, we would have a choice of taking the filter cutoff at either 50 or 200 and get one QTL either way. Under the weighted p-value approach, both QTLs are significant.

These examples show the tradeoffs in choosing different search schemes. Smaller search increments will tend to favor highly ranked features, because they can make a large proportional difference in the minimum number of tests including that feature. The chromosome 15 marker benefits greatly from being tested among 25 features rather than 100, even after correction factor. On the other hand, a lower ranked feature gets less proportional benefit and may decrease in discovery probability because of an increase in λ.

We have shown the results of several different search schemes in order to demonstrate important aspects of our method. However, it should be noted that the search scheme must be chosen in advance of seeing the data, or the Type I error rate will be inflated. Naïve adjustment of the search scheme to get the largest number of positive results will inflate false positives just as naïve adjustment of the filter cutoff does.

Table 2. The SNPs with the highest value of (max tests)/ (filter rank) for each of the four putative QTL regions

				Search scheme					
				100–500 by 100 $\lambda = 2.22$		25–1050 by 25 $\lambda = 4.17$		50–200 by 50 $\lambda = 2.03$	
ID	CHR.	Max tests	Filter rank	Next Highest Inc.	Corrected	Next Highest Inc.	Corrected	Next HighestInc.	Corrected
p45915	19	444	182	200	**444**	200	834	200	**406**
p44593	15	190	12	100	222	25	**104**	50	**102**
p45558	5	124	353	400	888	375	1564	–	–
p46339	1	53	101	200	444	125	521	150	305

4 Discussion

Although filtering methods have been in common use throughout the genomic era, their general statistical properties are not well understood. In this study, we have introduced a framework that quantifies filter effectiveness in terms of the probability of a feature associated with a true effect being ranked at or above a specified quantile. Using this approach, we have examined the conditions for filtering to be successful.

First, we have shown that filters can be very effective at increasing discovery probabilities for weak effects. However, the filter must have a substantial probability of ranking true positive features highly (e.g. the top 10%). Furthermore, the benefit of filtering is greater when there is high redundancy in information between hypothesis tests. In this case, a filter with a high probability of ranking true positive features in the upper 10% can increase discovery probabilities by 10-fold or more. The gain is less with lower redundancy between tests, but there is a several-fold (and higher) benefit over a wide range of conditions of filter effectiveness and redundancy in tests. In applications such as QTL mapping and GWAS, there are commonly many tests capable of detecting each true effect. In this case, filtering can be highly effective. In applications such as standard transcriptomics or proteomics where there is a single test per true effect, then redundancy is minimal and filtering less effective but can still be beneficial.

A major caveat is that the increased benefit in the case of high redundancy is contingent on the correlation in filtering ranks being low after conditioning on shared causative mechanism. If there is additional correlation (as in our QTL example where the ranks are all derived from the same data), this benefit is reduced (SI).

Second, the gain from filtering is highly dependent on the choice of filter cutoff. The choice of cutoff can easily make a 2- to 3-fold difference in discovery probability in the lower redundancy situation and a bigger difference in the high redundancy case. Most applications of filtering choose the cutoff in an arbitrary fashion, potentially leading to large loss in discovery probabilities relative to what is possible. Even worse, however, is choosing the cutoff based on the outcome without properly accounting for type I error. FWER can be greatly inflated when the cutoff is naively chosen based on the data.

Third, we have introduced a method for choosing the filter cutoff that finds the best value and properly accounts for the effect on FWER. Even after adjusting for the search procedure, there can still be a gain of several-fold in discovery probability relative to an arbitrary choice of cutoff. This data-dependent filtering procedure leads naturally to a closely related data-independent p-value weighting technique. In this approach, tests are filter-ordered in blocks and weighted by the inverse of the filter percentile rank for the lowest ranked member of the block, adjusted by a correction factor that ensures the target FWER is maintained. This method will always perform at least as well as optimal filtering and will sometimes result in more rejected null hypotheses while maintaining the same FWER.

The benefit of filtering depends strongly on effectiveness that the filter statistic is at ranking true effects highly. According to our results, a filter statistic with a high probability (say 70%) of ranking true effects in the top 10% can substantially improve discovery probabilities. Filters that are substantially less effective than this will not be likely to improve discovery probabilities and may make them worse. Determining whether filtering statistics can be expected to routinely perform this well is a crucial question that should be addressed in future research.

It is important to emphasize that *P*-value weighting/optimal filtering does come at a cost relative to the case where we know the correct filter cutoff in advance. A comparison of Figures 2 and 4 shows that discovery probabilities can be reduced by a factor of two or more between the case where we know the optimal filter cutoff and the case where we have to search for it. Furthermore, there is a tradeoff in deciding how finely we should search for the peak. Searching more finely makes it more likely to find the peak, but comes at a cost of a bigger FWER correction. When there is high correlation between hypothesis tests, then peaks tend to be sharper and searches with a finer grid may be beneficial. Future research should investigate the shape of filter functions for different types of data and filtering information. This would provide insight into the best choice of the vector $\vec{n}$. Ideally, we would determine where good cutoff points tend to be for particular types of data. Then, optimization of the filter would not be required and the full benefits of filtering could be realized.

Another finding from this work (Figure 1d and SI) is that filtering can greatly increase the relative discovery probability for weak effects, but it is constrained in terms of the absolute gains in power that are possible. Whether such gains for weak effects are of much value depends on the distribution of effect sizes. For example, accumulating evidence suggests that complex traits in humans are often driven by very large numbers of very low effect genetic variants (e.g. Gibson, 2011). In such a scenario, a boost in discovery probability from, for example, 0.1 to 1% for many such variants would be very significant. On the other hand, filters will be less effective at, for example, boosting power from 20% to 80% unless the filter is very effective at ranking true positive features highly.

A number of previous studies (Benjamini and Hochberg, 1997; Finos and Salmaso, 2007; Genovese, *et al.*, 2006; Holm, 1979; Ionita-Laza, *et al.*, 2007; Kropf, *et al.*, 2004; Roeder and Wasserman, 2009; Roquain and van de Wiel, 2009; Rubin, *et al.*, 2006; Wasserman, *et al.*, 2006; Westfall, *et al.*, 2004) have proposed alternative methods for *P*-value weighting and several studies (Roeder and Wasserman, 2009; Rubin, *et al.*, 2006; Wasserman, *et al.*, 2006) have derived optimal weights based on the true effect size. However, true effect size is never known and optimality is unclear under schemes for estimating it. Our correction factor approach is based solely on the filter-ranks. It will tend to perform better for sufficiently good filter-ranks, but the relative merits in practical circumstances are unclear. A future manuscript will explore these issues more thoroughly.

Funding

This study was supported in part by the University of Georgia Research Computing Center, a partnership between the Office of the Vice President for Research and the Office of the Chief Information Officer of the University of Georgia.

Conflict of Interest: none declared.

References

Benjamini,Y. and Hochberg,Y. (1997) Multiple hypotheses testing with weights. *Scand. J. Stat.*, **24**, 407–418.

Bourgon,R. (2010) Independent filtering increases detection power for high-throughput experiments. *Proc. Natl Acad. Sci.*, **107**, 9546–9551.

Bourgon,R. *et al.* (2010) Reply to Talloen et al.: independent filtering is a generic approach that needs domain specific adaptation. *Proc. Natl Acad. Sci.*, **107**, E175–E175.

Calle,M.L. *et al.* (2008) Improving strategies for detecting genetic patterns of disease susceptibility in association studies. *Stat. Med.*, **27**, 6532–6546.

Dai,J.Y. *et al.* (2012) Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika*, **99**, 929–944.

Degnan,J.H. *et al*. (2008) Genomics and genome-wide association studies: an integrative approach to expression QTL mapping. *Genomics*, **92**, 129–133.

Evans,D.M. *et al*. (2006) Two-stage two-locus models in genome-wide association. *PLoS Genet*, **2**, e157.

Finos,L. and Salmaso,L. (2007) FDR- and FWE-controlling methods using data-driven weights. *J. Stat. Plan. Inference*, **137**, 3859–3870.

Fog,A. (2008a) Calculation methods for Wallenius' noncentral hypergeometric distribution. *Commun. Stat. Simul. C*, **37**, 258–273.

Fog,A. (2008b) Sampling methods for Wallenius' and Fisher's noncentral hypergeometric distributions. *Commun. Stat. Simul. C*, **37**, 241–257.

Genovese,C.R. *et al*. (2006) False discovery control with p-value weighting. *Biometrika*, **93**, 509–524.

Ghazalpour,A. *et al*. (2006) Integrating genetic and network analysis to characterize genes related to mouse weight. *Plos Genet*. **2**, 1182–1192.

Gibson,G. (2011) Rare and common variants: twenty arguments. *Nat. Rev. Genet*., **13**, 135–145.

Hackstadt,A.J. and Hess,A.M. (2009) Filtering for increased power for microarray data analysis. *BMC Bioinf*., **10**, 11.

Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat*., **6**, 65–70.

Ionita-Laza,I. *et al*. (2007) Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. *Am. J. Hum. Genet*., **81**, 607–614.

Jiang,H. and Doerge,R.W. (2006) A two-step multiple comparison procedure for a large number of tests and multiple treatments. *Stat. Appl. Genet. Mol. Biol*., **5**, Article28.

Kropf,S. *et al*. (2004) Nonparametric multiple test procedures with data-driven order of hypotheses and with weighted hypotheses. *J. Stat. Plan. Inference*, **125**, 31–47.

Li,L. *et al*. (2013) Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Front. Genet*., **4**, 103.

Lu,J. *et al*. (2011) Principal component analysis-based filtering improves detection for Affymetrix gene expression arrays. *Nucleic Acids Res*., **39**, e86.

McClintick,J.N. and Edenberg,H.J. (2006) Effects of filtering by Present call on analysis of microarray experiments. *BMC Bioinf*., **7**, 49.

Pattin,K.A. and Moore,J.H. (2008) Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. *Hum. Genet*., **124**, 19–29.

Patwardhan,A. *et al*. (2014) Variant priorization and analysis incorporati-ng problematic regions of the genome. *Pac. Symp. Biocomput*., 277–287.

Ramskold,D. *et al*. (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol*., **5**, e1000598.

Rau,A. *et al*. (2013a) HTSFilter : independent data-based filtering for replicated transcriptome sequencing experiments. 1–14, web document found at https://www.bioconductor.org/packages/release/bioc/vignettes/HTSFilter/inst/doc/HTSFilter.pdf.

Rau,A. *et al*. (2013b) Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics*, **29**, 2146–2152.

Roeder,K. and Wasserman,L. (2009) Genome-wide significance levels and weighted hypothesis testing. *Stat. Sci. Rev. J. Inst. Math. Stat*., **24**, 398–413.

Roquain,E. and van de Wiel,M.A. (2009) Optimal weighting for false discovery rate control. *Electron. J. Stat*., **3**, 678–711.

Rubin,D. *et al*. (2006) A method to increase the power of multiple testing procedures through sample splitting. *Stat. Appl. Genet. Mol. Biol*., **5**, 19.

Smith,C.M. *et al*. (2014) The mouse Gene Expression Database (GXD): 2014 update. *Nucleic Acids Res*., **42**, D818–D824.

Sultan,M. *et al*. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science (New York, N.Y)*, **321**, 956–960.

Talloen,W. *et al*. (2007) I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics (Oxford, England)*, **23**, 2897–2902.

Talloen,W. *et al*. (2010) Filtering data from high-throughput experiments based on measurement reliability. *Proc. Natl Acad. Sci. USA*, **107**, E173–E174.

Wasserman,L. *et al*. (2006) Genome-wide significance levels and weighted hypothesis testing. *Stat. Sci. 2009*, **24**, 398–411.

Westfall,P.H. *et al*. (2004) Weighted FWE-controlling methods in high-dimensional situations. *Lect. Notes Monogr. Ser. Recent Dev. Multiple Comparison Proced*., **47**, 143–154.