Optimal Multiple Testing Under a Gaussian Prior on the Effect Sizes

Edgar Dobriban*[1], Kristen Fortney[2], Stuart K. Kim[2] and Art B. Owen†[1]

[1]Department of Statistics, Stanford University, Stanford, California 94305, U.S.A.
[2]Department of Developmental Biology, and Department of Genetics, Stanford University, Stanford, California 94305, U.S.A.

April 11, 2015

Abstract

We develop a new method for frequentist multiple testing with Bayesian prior information. Our procedure finds a new set of optimal p-value weights called the Bayes weights. Prior information is relevant to many multiple testing problems. Existing methods assume fixed, known effect sizes available from previous studies. However, the case of uncertain information is usually the norm. For a Gaussian prior on effect sizes, we show that finding the optimal weights is a non-convex problem. Despite the non-convexity, we give an efficient algorithm that solves this problem nearly exactly. We show that our method can discover new loci in genome-wide association studies. On several data sets it compares favorably to other methods. Open source code is available.

Contents

1 Introduction **2**

2 Related Work **3**

3 Theoretical results **4**
3.1 Background 4
3.2 Weighting leads to a substantial power gain 5
3.3 Weights with imperfect prior knowledge 5

4 Simulation studies **8**
4.1 Bayes weights are more powerful than competing methods 8
4.2 Bayes weights have a worst-case advantage 9

5 Application to Genome-Wide Association Studies **10**
5.1 Review of GWAS 10
5.2 Prior Information 11
5.3 Practical remarks 11

*corresponding author: dobriban@stanford.edu
†owen@stanford.edu

6 Data Analysis **12**
6.1 Data Sources . . . 12
6.2 Specific Pairs . . . 12
6.3 Methods compared . . . 12
6.4 Additional details . . . 13
6.5 Results . . . 13

7 Acknowledgements **14**

8 Software **14**
8.1 End-user software . . . 14
8.2 Reproducibility . . . 14

9 Proofs **14**
9.1 Proof of Proposition 1 - Sparse Means . . . 14
9.2 Proof of Theorem 1 - Optimal weights for small q . . . 15
9.3 Proof of Proposition 2 - Recovering the Spjotvoll weights . . . 16
9.4 Proof of Proposition 3 - Simple sufficient condition . . . 17
9.5 Proof of Theorem 2 - Optimal weights in the general case . . . 17
9.5.1 Monotonicity of the generic term . . . 17
9.5.2 Method for computing weights . . . 19

10 Data sources **21**
10.1 CARDIoGRAM - CAD . . . 21
10.2 C4D - CAD . . . 21
10.3 Chronic Kidney Disease Consortium (CKDGen) - eGFR creatinine . . . 22
10.4 Blood Lipids . . . 22
10.5 Psychiatric Genomics Consortium - Schizophrenia . . . 22

1 Introduction

We are motivated by the genetics of human longevity. Genome-wide association studies of longevity compare long-lived individuals (e.g. centenarians who live to 100 or older) to matched controls (Brooks-Wilson, 2013). More than 500,000 genetic variants are tested for association to longevity. This is a large multiple testing problem. In addition to the multiplicity, the sample size is low - usually a few hundred. As a consequence, only a few loci have been replicably associated to human longevity. They do not explain the heritability of the trait (Hjelmborg et al., 2006).

The multiplicity may be countered by testing only a few candidate variants selected based on prior scientific knowledge. In a separate work in preparation, led by Dr. Kristen Fortney, we find that a more general genome-wide test helps improve power in longevity. We leverage prior information from genome-wide association studies of age-related diseases, such as coronary artery disease and diabetes. For this task, we develop a new method of frequentist multiple testing with Bayesian prior information. In this paper we provide the theory for this method.

Our method is a type of p-value weighting scheme. P-value weighting is a general methodology for multiple testing that leverages independent prior information to improve power (see Roeder and Wasserman (2009); Gui et al. (2012) for a review). Suppose we test hypotheses H_i $i = 1, 2, \ldots J$, via the p-values P_i. For a significance level $q \in [0, 1]$, the weighted Bonferroni method declares the

i-th hypothesis significant if $P_i \leq qw_i$. The weights $w_i \geq 0$ are based on independent data. For weights averaging to 1, the family-wise error rate (the probability of making at least one error) is controlled strongly at $\alpha = Jq$.

Previous work has found the optimal weights in a Gaussian model of hypothesis testing. Let the test statistics in the current study be $T_i \sim \mathcal{N}(\mu_i, 1)$, where μ_i are the *means*, or *effect sizes*, and test the null hypotheses $\mu_i \geq 0$ against $\mu_i < 0$. We have some information about μ_i from prior studies. The works Wasserman and Roeder (2006); Roeder and Wasserman (2009); Rubin et al. (2006) consider the model where μ_i is known *exactly* from the prior data, and the weights are allowed to depend on μ_i. In this model they find the optimal weights maximizing the expected number of discoveries. We show this amounts to solving a convex optimization problem, but this was not used originally.

The assumption that μ_i are known precisely is problematic: if they were known, there should be no follow-up study. Instead, we account for uncertainty by considering the model $\mu_i \sim \mathcal{N}(\eta_i, \sigma_i^2)$. Here only the prior mean η_i and standard error σ_i are known from independent data, not the precise value of the effect sizes. However, finding the optimal weights, which we call Bayes weights, becomes a non-convex optimization problem. Westfall et al. (1998) use a direct numerical solver, and optimize for at most 4 tests.

We give an efficient method to find the optimal weights for a large number of tests. We solve the optimization problem exactly for small q. For larger q, we can solve it for a nearby q^* such that $|q^* - q| \leq 1/(2J)$. The cost in the first case is $O(J)$, the cost in the second case is $O(J \log J)$. These are the costs per iteration of our optimization algorithm. We observe a near constant number of iterations used. The solution of non-convex optimization problems is challenging in general, thus it is perhaps remarkable that this problem admits a nearly exact solution.

This enables a new methodology for large-scale multiple testing that controls a frequentist error measure, while also taking into account Bayesian prior information. This method follows George Box's advice to be Bayesian when predicting but frequentist when testing (Box, 1980). While this methodology was considered previously for a few tests (Westfall et al., 1998), we are the first to do it on a large scale.

When prior information is uncertain, we show in simulations that the new scheme does better than its competitors. We also show theoretically, in a sparse mixture model, that weighting leads to substantially improved power. We apply our method to genome-wide association studies (GWAS). By analyzing several GWAS data sets we show its advantages compared to other methods.

This method should be useful for other problems in biology and elsewhere. Open source code is available from the authors. All our computational results are reproducible (see the Supplement).

The contents of the paper are as follows: We discuss related work in Section 2. We develop the theory in Section 3 and present simulations comparing our method to alternatives in Section 4. We apply our method to genome-wide association studies in Section 5 and use it to analyze GWAS data in Section 6. The supplementary material contains a description of available software and instructions to reproduce our computational results (Section 8), as well as mathematical proofs (Section 9).

2 Related Work

There is a large literature on related statistical methods for multiple testing with prior information. In early work, Spjotvoll (1972) devised optimal multiple testing procedures treating tests unequally. Later it was recognized that Spjotvoll's results are equivalent to optimal p-value weighting methods. For instance, Benjamini and Hochberg (1997) developed extensions of Spjotvoll's methods for p-

value weighting.

Leveraging Spjotvoll's results, Wasserman and Roeder (2006); Roeder and Wasserman (2009); Rubin et al. (2006) found the explicit formula for optimal weights in the Gaussian model $\mathcal{N}(\mu_i, 1)$, assuming the effects are known exactly. This lead to an efficient method suitable for large applications. In the bioinformatics community Eskin (2008); Darnell et al. (2012) applied Wasserman and Roeder (2006)'s framework to GWAS. They accounted for correlations between the tests but assumed the effects are known exactly.

The simple method of testing the top candidates from a prior study is also popular. This is often known as two-stage testing or as a candidate study. A specific version for GWAS has been called the proxy-phenotype method (Rietveld et al., 2014). Using only the top candidates runs the risk of discarding many potentially useful hypotheses.

A missing ingredient is taking uncertainty into account. Westfall et al. (1998) considered a Gaussian model $\mathcal{N}(\mu_i, 1)$ for hypothesis testing where prior distributions are known for the means. However, their optimization methods could handle only a small number ($J = 4$) of tests.

3 Theoretical results

3.1 Background

In this section we present our theoretical results. As background, we begin with the case of known means. We work in the Gaussian means model of hypothesis testing: We observe test statistics $T_i \sim \mathcal{N}(\mu_i, 1)$ and test each null hypothesis $H_i : \mu_i \geq 0$ against $\mu_i < 0$. The p-value for testing H_i is $P_i = \Phi(T_i)$, where Φ is the normal cumulative distribution function.

For a weight vector $w \in [0, \infty)^J$ and significance level $q \in [0, 1]$, the weighted Bonferroni procedure rejects H_i if $P_i \leq qw_i$. Usual Bonferroni corresponds to $w_i = 1$. Then the expected number of false rejections, known as the per-family error rate, equals: $\sum_{\mu_i \geq 0} \mathrm{pr}(P_i \leq qw_i) = q \sum_{\mu_i \geq 0} w_i$. Therefore, if $\sum_{i=1}^J w_i \leq J$, the expected number of false rejections is controlled strongly, under any configuration of truth or falsehood of H_i, at $\alpha := Jq$. By Markov's inequality this implies that the family-wise error rate, the probability of any false rejection, is also controlled at α. This does not need independence of the T_i. We always assume that $q \leq 1$, and usually $q \ll 1$. Without loss of generality we restrict the weights to $[0, 1/q]$.

Let us denote the number of rejections by $R(w) = \sum_{i=1}^J I(P_i \leq qw_i)$, where $I(\cdot)$ is the indicator function. The optimal weights in this model were found explicitly by Wasserman and Roeder (2006); Roeder and Wasserman (2009) and independently by Rubin et al. (2006). Denoting by $E_T(\cdot)$ expectation with respect to T_i, they solve the constrained optimization problem

$$\max_{w \in [0,1/q]^J} E_T\{R(w)\} \text{ subject to } \sum_{i=1}^J w_i = J. \tag{1}$$

It was not noted in these works that this problem is convex. Usually convex programs are about minimization of convex functions. This is equivalent to maximizing concave functions. In our case the objective is a sum of terms of the form $\Phi\{\Phi^{-1}(qw_i) - \mu_i\}$, whose concavity follows directly by differentiation. Yet, by simple Langrangian optimization, the above papers show that if all $\mu_i < 0$, the optimal weights are $w_i = w(\mu_i)$, where

$$w(\mu) = \Phi\left(\frac{\mu}{2} + \frac{c}{\mu}\right)/q. \tag{2}$$

Here c is the unique normalizing constant such that the weights sum to J. Interestingly, the weights are not monotonic as a function of μ, but maximal for intermediate values of μ (Roeder and Wasserman, 2009).

As noted by Roeder and Wasserman (2009) the formula is a direct consequence of Spjotvoll's theory of optimality in multiple testing (Spjotvoll, 1972). Accordingly, we call these the *Spjotvoll weights*.

3.2 Weighting leads to a substantial power gain

To illustrate theoretically that weighting can lead to increased power, we compare the power of optimal weighting and unweighted testing in a sparse mixture model.

P-value weighting exploits the heterogeneity of the tests. In the simplest case there are only large and small effects, say $M \ll m \approx 0$. We want the $m \to 0$ limit, and for simplicity we suppose $m = 0$. Similar results hold if $m \approx 0$. Let the fraction of large and small means be $\pi_1, \pi_0 > 0$, so that $\pi_1 J$ means are M, and the remaining $\pi_0 J$ are 0. We solve below for the optimal weights.

Proposition 3.1 (Optimal weights for sparse means)**.** *There is a set of optimal weights that gives the same weight to equal means, w_0 and w_1 to 0 and M. These are:*

$$(w_0, w_1) = \begin{cases} (0, 1/\pi_1) & \text{if} \quad \pi_1\Phi(-|M|/2) > q, \\ \left\{\frac{q-\pi_1\Phi(-|M|/2)}{q\pi_0}, \frac{\Phi(-|M|/2)}{q}\right\} & \text{if} \quad \pi_1\Phi(-|M|/2) \le q. \end{cases}$$

Further, the power p^* of the optimal p-value weighting method is:

$$p^*(\pi_1, M, q) = \begin{cases} \pi_1\Phi\left\{\Phi^{-1}(q/\pi_1) + |M|\right\} & \text{if} \quad \pi_1\Phi(-|M|/2) > q, \\ q + \pi_1\left\{\Phi(|M|/2) - \Phi(-|M|/2)\right\} & \text{if} \quad \pi_1\Phi(-|M|/2) \le q. \end{cases}$$

If $|M|$ is small enough that $\pi_1\Phi(-|M|/2) > q$, all the weight is placed on the larger means. This is the behavior we expect intuitively. However, if $|M|$ is large enough that $\pi_1\Phi(-|M|/2) \le q$, then it is advantageous to place some weight on the small means. The reason is that such a large $|M|$ will be detected with very high probability.

The power of unweighted Bonferroni is $p_{\text{unif}}(\pi_1, M, q) = \pi_0 q + \pi_1\Phi\{\Phi^{-1}(q) + |M|\}$.

In Figure 1 (a), we plot the ratio of $p^*(\pi_1, M, q)/p_{\text{unif}}(\pi_1, M, q)$ for $q = 10^{-3}$ for a range of M and π_1. We observe that for most effect sizes $M \in [-2.5, -0.25]$, and $\pi_1 < 0.4$, we get a power boost of at least 50%. Further, there is a hotspot where the power gain can be 3-4 fold. Thus, optimal weighting can lead to a significant boost in power.

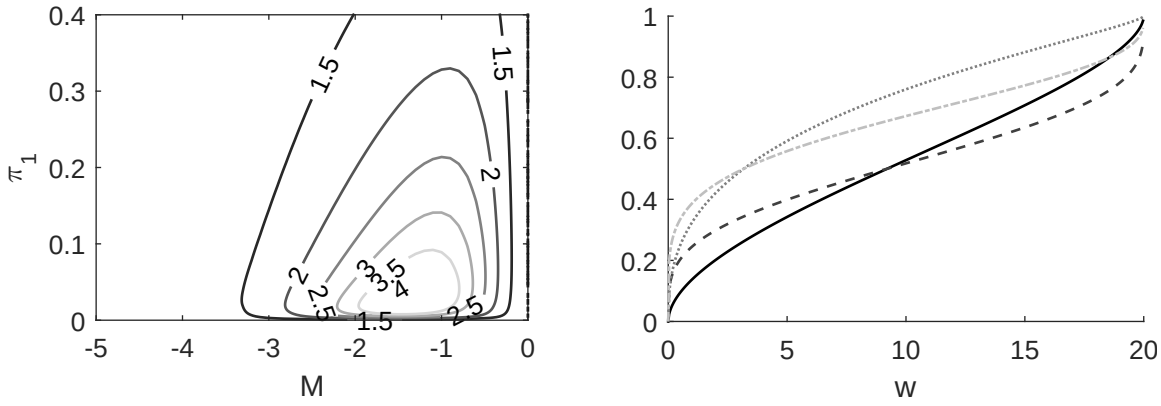
3.3 Weights with imperfect prior knowledge

In the previous sections it was assumed that the effects μ_i were known precisely. We will instead assume that we have uncertain prior information $\mu_i \sim \mathcal{N}(\eta_i, \sigma_i^2)$ about them.

We maximize the expected power $E_\mu[E_T\{R(w)\}]$. The expectation is with respect to the random T_i and μ_i. Introducing $\gamma_i = (\sigma_i^2 + 1)^{1/2}$, the *Bayes weights* problem becomes:

$$\max_{w\in[0,1/q]^J} \sum_{i=1}^{J} \Phi\left\{\frac{\Phi^{-1}(qw_i) - \eta_i}{\gamma_i}\right\} \quad \text{subject to} \quad \sum_{i=1}^{J} w_i = J. \tag{3}$$

This objective is not concave if any $\gamma_i > 1$. To help with visualization, the function $w \to \Phi[\{\Phi^{-1}(qw) - \eta\}/\gamma]$ is plotted in Figure 1 (b) for four parameter pairs (η, γ). The function is increasing, and its curvature has two intervals: concavity, followed by convexity.



(a) Power gain.

(b) Non-convex summand.

Figure 1: (a) Contour plot of power ratio of optimal and unweighted testing for sparse means. (b) Plots of four different instances of the function that is summed in the optimization objective. The non-convex function $w \to \Phi[\{\Phi^{-1}(qw) - \eta\}/\gamma]$ with $q = 0.05$ is plotted for the following pairs (η, σ): $(-0 \cdot 1, 1)$ solid, $(-0 \cdot 1, 2)$ dashed, $(-1, 1)$ dotted, $(-1, 2)$ dot-dashed.

Our key contribution is to solve this problem efficiently for large J. Our results here are twofold. First, we can solve the problem exactly in the special case when q is sufficiently small. Second, we have a nearly exact solution for arbitrary q. We start with the simpler first case. Let us define

$$c(\eta, \gamma; \lambda) = -\frac{\eta + \gamma\{\eta^2 + 2\left(\gamma^2 - 1\right)\log(\gamma\lambda)\}^{1/2}}{\gamma^2 - 1}. \tag{4}$$

It turns out that c is the optimal critical value when

$$q \leq \frac{1}{J}\sum_{i=1}^{J} \Phi\left\{c(\eta_i, \gamma_i; 1)\right\}. \tag{5}$$

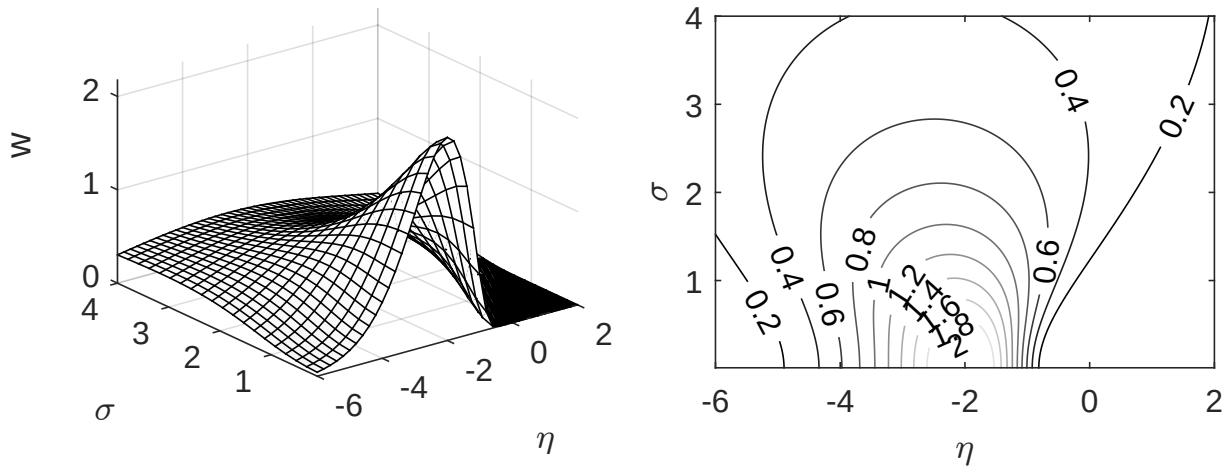
In our data analysis examples and simulations, this upper bound requires that q be below values in the range $0 \cdot 1 - 0 \cdot 3$. In the next result we find the exact optimal weights for small q when all $\sigma_i > 0$.

Theorem 3.2 (Form of Bayes weights)**.** *If the significance level $q \in [0, 1]$ is small enough that (5) holds then the optimal* Bayes weights *maximizing the average power (3) are $w_i = w(\eta_i, \gamma_i; \lambda) = \Phi\{c(\eta_i, \gamma_i; \lambda)\}/q$, where $\lambda \geq 1$ is the unique constant such that $\sum_{i=1}^{J} w(\eta_i, \gamma_i; \lambda) = J$.*

In the supplementary material, we solve this problem by maximizing the Lagrangian. The two key properties are the joint separability of the objective and constraint; and the analytic tractability of the Gaussian.

Figure 2 shows surface and contour plots of an instance of the optimal weights $w(\eta, \sigma)$, as a function of the mean η and standard deviation σ. In the theorem the weights are a function of (η, γ). Here and below, we will often view them as a function of (η, σ), via the natural map $\gamma^2 = \sigma^2 + 1$. As the standard error σ becomes small, our weights tend to the Spjotvoll weights:

Proposition 3.3 (Recovering the Spjotvoll weights)**.** *For any λ and $\eta < 0$, the Bayes weight function defined by $w(\eta, \gamma; \lambda) = \Phi\{c(\eta, \gamma; \lambda)\}/q$ tends to the Spjotvoll weight defined in (2) as $\sigma \to 0$.*



(a) Surface

(b) Contour

Figure 2: Bayes weights: (a) surface and (b) contour plots of the Bayes weight function $w(\eta, \sigma)$ defined in Theorem 3.2. Spjotvoll weights are on the segment $\sigma = 0$, $\eta < 0$.

With $\sigma_i > 0$, the weights are regularized: more extreme weights are shrunk towards a common value in a nonlinear way. For finite σ_i our weights can be viewed as a smooth interpolation between Spjotvoll and uniform weights. It is reasonable to think at first that as all $\sigma_i \to \infty$, the best weight allocation is uniform. This is not the case. As we will show below, an interesting symmetry breaking phenomenon occurs.

Consider a weight vector w that equals $1/q$ for $\lfloor Jq \rfloor$ indices, and assume Jq is not an integer. Distribute the remaining strictly positive weight equally among the remaining hypotheses. Now it is easy to see that for the $1/q$ weights we always reject, thus the power equals 1. For the remaining weights the objective $\Phi[\{\Phi^{-1}(qw) - \eta\}/\gamma]$ tends to $\Phi(0) = 1/2$ as $\sigma \to \infty$ (i.e. $\gamma \to \infty$). This shows that the limiting power is $\lfloor Jq \rfloor + (J - \lfloor Jq \rfloor)/2 = (J + \lfloor Jq \rfloor)/2$. This is larger than $J/2$, the limit power of uniform weighting! This illustrates the symmetry breaking phenomenon caused by the extreme non-convexity of the optimization problem.

Fortunately the situation is better as long as condition (5) holds. This condition is easy to check for any given parameters q and (η_i, σ_i), $i = 1, \ldots, J$. In addition we will now show theoretically that the constraint is mild. Often, even if J is large, we want to keep $\alpha = Jq$ small, because α is the number of false rejections we tolerate. In this regime the condition holds as long as there are a few average-sized negative prior means η_i (denote $z_c = \Phi^{-1}(c)$).

Proposition 3.4 (Simple condition)**.** *Condition* (5) *is true (and so Theorem 3.2 applies) if there are K distinct indices i with negative η_i, such that $\gamma_i^2 \log(\gamma_i^2)/|z_{\alpha/K}| \le |\eta_i| \le |z_{\alpha/K}|$.*

If $\alpha/K \to 0$, it is well-known that asymptotically $|z_{\alpha/K}| \sim \{2\log(K/\alpha)\}^{1/2}$, so the simple condition holds as long as: $\gamma_i^2 \log(\gamma_i^2)\{2\log(K/\alpha)\}^{-1/2} \lesssim |\eta_i| \lesssim \{2\log(K/\alpha)\}^{1/2}$. This is a quite weak requirement. For instance, if $K = 10$, $\alpha = 0 \cdot 01$, then $\{2\log(K/\alpha)\}^{1/2}$ approximately equals 3.7. Continuing this example, if $\sigma = 1$, so that $\gamma^2 = 2$, and $\gamma^2 \log(\gamma^2)/3.7 = 0.16$, then we only need 10 hypotheses with $0.16 \le |\eta| \le 3.7$.

When q is small, we use Newton's method to find the right constant λ from the theorem via a one-dimensional line search. The function evaluations cost $O(J)$ per iteration, and empirically it takes a small number of iterations independent of J to converge. In our data analysis section, we solve problems with more than 2 million tests in a few seconds on a desktop computer.

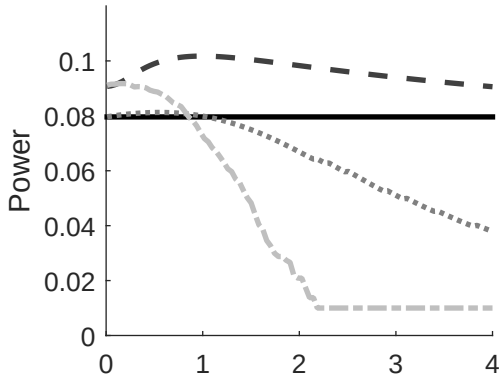


Figure 3: Power of four p-value weighting methods as a function of their parameter. Unweighted (solid), Bayes (dashed) as a function of the dispersion ϕ, exponential (dotted) as a function of β, and filtering (dot-dashed) as a function of $|M|$. The Spjotvoll weights correspond to the point at the origin $\phi = 0$ on the Bayes weights curve.

Now we move to presenting our result for the general case.

Theorem 3.5 (Weights in the general case)**.** *For any $q \in [0, 1]$, the non-convex Bayes weights problem can be solved for a nearby $q^* \in [0, 1]$, for which $|q^* - q| \leq 1/(2J)$. The optimal weights and q^* can be found in $O(J \log J)$ steps.*

This result is most relevant for the settings when $\alpha = Jq$, the expected number of errors under the null, is set to at least $1/2$. If so, and especially for large J, our weights will be optimal for a q^* that is close to q. We see from the proof that even for large q, q^* often equals q. The method also returns the value q^*, which the user can inspect. It is then up to the user to choose whether to perform multiple testing adjustment at the original level q or at the new level q^*.

We note that the analysis of non-convex optimization problems is challenging in general. It is perhaps remarkable that the non-convex Bayes weights problem admits a nearly exact solution.

4 Simulation studies

4.1 Bayes weights are more powerful than competing methods

We present two simulation studies to explore the empirical performance of our method. First we show that Bayes weights increase power more reliably than competing methods. We compare three methods of p-value weighting: Bayes, exponential, and filtering.

For Bayes weights we multiply the variances by a dispersion factor ϕ: $\mathcal{N}(\eta_i, \phi\sigma_i^2)$. By changing the dispersion, we test the robustness of our method to mis-specification of the prior variances. This corresponds to the same dispersion variable ϕ in our GWAS application given in the next section. The dispersion ranges from 0 to 4. Spjotvoll weights correspond to $\phi = 0$.

Exponential weights with tilt β are defined as: $w_i = \exp(\beta|\eta_i|)/c$, where $c = \sum_{i=1}^{J} \exp(\beta|\eta_i|)$. This weighting scheme was proposed in (Roeder et al., 2006), who recommend $\beta = 2$ as a default. We include the range $\beta \in [0, 4]$. As noted by Roeder et al. (2006), exponential weights are sensitive to large means. To guard against this sensitivity, we truncate the weights larger than $1/q$ and re-distribute their excess weight among the next largest weights.

Filtering methods test only the most significant effects $\eta_i \leq M$, and give them equal weights. This and related methods are known under many names, such as two-stage testing, screening, or

the proxy phenotype method (Rietveld et al., 2014). We adopt the name filtering from Bourgon et al. (2010), who filter based on independent information in the current data set instead of prior information. The threshold M ranges from -4 to 0. If $|M|$ is large and fewer than Jq hypotheses would be tested, then we instead test the most significant Jq.

The simulation is conducted as follows: We generate $J = 1000$ random means and variances independently according to $\eta_i \sim \mathcal{N}(0,1)$, $\sigma_i \sim |\mathcal{N}(0,1)|$. We set $q = 10^{-2}$. For any weight vector w, we calculate the power as the objective from (3) divided by J, to reflect the average power per test.

The results are shown in Figure 3. Each method can improve the power over unweighted testing. However, Bayes weights lead to more power than the other methods. The best power is attained when the dispersion $\phi = 1$, but good power is reached even when ϕ is not 1. Our weights are robust to mis-specifying the dispersion.

Note, in particular, that taking uncertainty into account helps. Spjotvoll weights, which assume fixed and known effects, and are shown on the figure as regularized weights with $\phi = 0$, have less power than Bayes weights with positive ϕ, for a wide range of ϕ.

The remaining two methods, filtering and exponential weights, have disadvantages. While filtering leads to power gain for a thresholding parameter $M \lesssim 3/4$, it also leads to a substantial power loss for $M > 1$. For sufficiently large M the power equals q, because only the top Jq hypotheses are selected. Further, it is a significant disadvantage that there is no principled way to choose M a priori without additional assumptions.

Similarly, exponential weighting leads to at most a small gain in power, and usually leads to a loss. There also appears to be no simple, principled way to choose β a priori.

We conclude that Bayes weights are quite insensitive to tuning and have uniformly good power. In contrast, exponential weighting and filtering are relatively sensitive and their power can drop substantially. Therefore, Bayes weights increase power more reliably than competing methods.

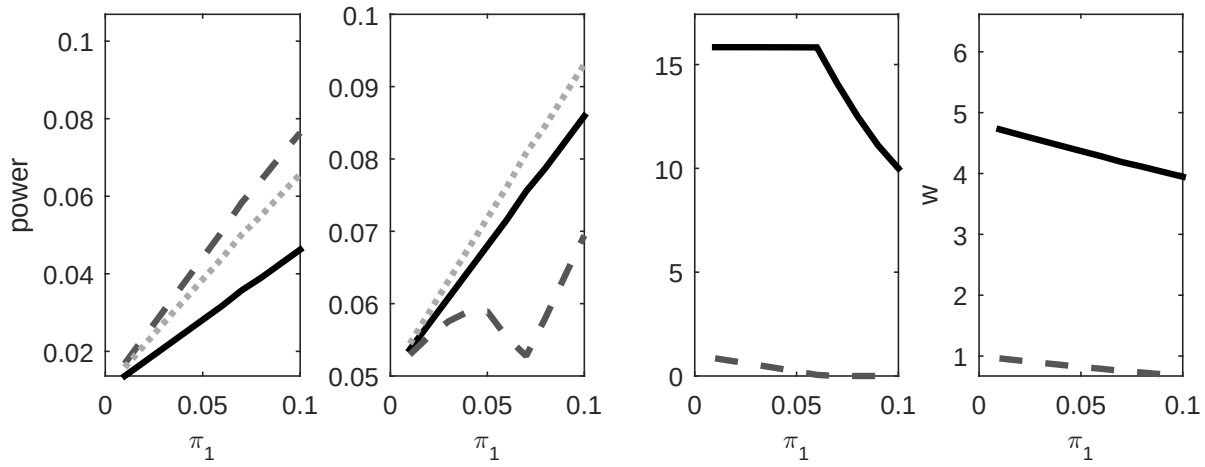
4.2 Bayes weights have a worst-case advantage

We show that Bayes weights have a worst-case advantage compared to Spjotvoll weights. We use the sparse means model: we generate $J = 1000$ means η_i, such that their distribution is $\eta \sim \pi_0\delta_m + \pi_1\delta_M$, where $m = -10^{-3}$ and $M = -2$. We set $q = 10^{-2}$ and we vary π_1 from 0 to 0.1. We set all $\sigma_i = \sigma$.

We consider two values of σ: 0 and 1. Spjotvoll is optimal for $\sigma = 0$, while Bayes weights with $\sigma = 1$ are optimal for 1. We evaluate these weighting schemes by calculating the objective *that they do not maximize*: the average power (3) for Spjotvoll and the deterministic power (1) for Bayes weighting. We also compute the power of the unweighted Bonferroni method.

The results are in Figure 4. Bayes weights lose only a little compared to the optimal Spjotvoll ((a) left). In contrast, Spjotvoll loses a lot compared to Bayes ((a) right). Bayes maximizes the worst-case power, thus showing a maximin property.

Spjotvoll weights show a marked drop in power near $\pi_1 = 0.07$, as shown by its non-monotonic power curve in Figure 4 (a)). To understand this, we plot the two weighting schemes in Figure 4 (b). Since there are only two classes, the weights also take two values. We see that the Spjotvoll weights are more extreme than the Bayes ones. They start putting weight *equal to zero* on the small means near $\pi_1 = 0.07$, which appears to lead to power loss.



(a) Power comparison for sparse means.

(b) Weights of the two classes.

Figure 4: (a) Deterministic (left) and average (right) power as a function of the proportion of large means π_1. Three methods are compared: unweighted (solid), Spjotvoll (dashed), Bayes (dotted). (b) The weights of the two classes, large (solid) and small (dashed), for Spjotvoll (left) and Bayes (right), as a function of the proportion of large means π_1.

5 Application to Genome-Wide Association Studies

5.1 Review of GWAS

We adapt our framework to genome-wide association studies, relying on basic notions of quantitative genetics (see e.g. Lynch and Walsh, 1998). Our method applied to this problem is called iGWAS in our forthcoming application to human longevity. This section presents in detail the methodology for that application, while also illustrating the steps to use our framework for specific problems.

Consider a model for GWAS in which we study a quantitative trait y in a population. Our goal is to understand the effects of single nucleotide polymorphisms (SNPs) $g_1, g_2, \ldots, g_J$ on the trait. We assume y has mean 0 and known variance, and g_i denotes the centered minor allele count of variant i for an individual. We rely on the linear model for the effect of the i-th variant on the trait: $y = g_i\beta_i + \varepsilon_i$.

We will show how our question can be framed in the Gaussian means model of hypothesis testing. In the above model y is the phenotype of a randomly sampled individual from the population. Accordingly, g_i is random, β_i is a fixed unknown constant, and ε_i is the residual error. This error is a mean-zero random variable independent of g_i, with variance σ_i^2.

Suppose we observe a sample of N independent and identically distributed observations from this model. We use the standard linear regression estimate $\hat{\beta}_i$, which for large sample size has an approximate distribution $N^{1/2}\hat{\beta}_i \dot{\sim} \mathcal{N}\{N^{1/2}\beta_i, \sigma_i^2/\mathrm{var}(g_i)\}$. We can standardize if we divide by τ_i, where $\tau_i^2 = \sigma_i^2/\mathrm{var}(g_i)$ is the variance of $N^{1/2}\hat{\beta}_i$.

With these steps, we have framed our problem in the Gaussian means model. Denoting $T_i = N^{1/2}\hat{\beta}_i/\tau_i$, $\mu_i = N^{1/2}\beta_i/\tau_i$, we have $T_i \dot{\sim} \mathcal{N}(\mu_i, 1)$, which is the required form. Denote also the standardized effect size $\nu_i = \beta_i/\tau_i$, which will be of key importance.

5.2 Prior Information

Now we show how to use prior information. Assume we also have a *prior* trait y_0 measured independently on a different, independent sample from the same population. If our assumptions also hold for y_0, we can write $y_0 = g_i\beta_{0i} + \varepsilon_{0i}$. Here β_{0i} is a fixed unknown constant, and ε_{0i} is random. Suppose we have independent samples of size N_i and N_{0i} for the two traits. If we define T_{0i}, ν_{0i} by analogy to their definitions for y, we can write $T_{0i} \sim \mathcal{N}(N_{0i}^{1/2}\nu_{0i}, 1)$.

We model the relatedness between the two traits as a relation between the standardized effect sizes ν, which do not depend on the sample size. If the two traits are closely related, the first order approximation is equality: $\nu_i = \nu_{0i}$. This simple model captures the pleiotropy between the two traits (e.g. Solovieff et al., 2013).

The final step is to compute the distribution of ν_i given the prior data T_{0i}. For this we need to choose a prior for ν_i, and for simplicity we will use a flat prior.

We now have all ingredients to apply the model for Gaussian hypothesis testing with uncertain information. Specifically, we have $\mu_i \sim \mathcal{N}(\eta_i, \sigma_i^2)$, where $\mu_i = N_i^{1/2}\nu_i$, $\eta_i = (N_i/N_{0i})^{1/2}T_{0i}$, and $\sigma_i^2 = N_i/N_{0i}$.

The uncertainty in T_{0i} may be different from 1, for instance larger than 1 due to overdispersion. In addition, overdispersion is one way to weaken the first order approximation assumption. To allow for this, we recall the dispersion parameter ϕ used in our simulation. We model the prior data as $T_{0i} \sim \mathcal{N}(N_{0i}^{1/2}\nu_{0i}, \phi)$. Then the variance $\sigma_i^2 = \phi N_i/N_{0i}$. The default is $\phi = 1$. Finally, we compute Bayes weights w_i with parameters q, (η_i, σ_i^2), and run weighted Bonferroni on the current p-values. This fully specifies the method. For the reader's convenience, the method is summarized in Algorithm (1).

Algorithm 1 Multiple Testing with Bayes weights in GWAS

$T_{0i} \leftarrow$ prior effect sizes for $i = 1, \ldots, J$
$N_{0i}, N_i \leftarrow$ prior and current sample sizes
$P_i \leftarrow$ current p-values
$q \leftarrow$ significance threshold
$\phi \leftarrow$ dispersion (default $\phi = 1$)
Set the prior means and variances: $\eta_i = (N_i/N_{0i})^{1/2}T_{0i}, \quad \sigma_i^2 = \phi N_i/N_{0i}$
Compute Bayes weights w_i, defined via (3), with parameters q, (η_i, σ_i^2)
Return indices i such that $P_i \leq qw_i$

5.3 Practical remarks

It is important that we retain type I error control as soon as we have valid p-values, even if the modelling assumptions fail. Common deviations from our model are: **(1)** GWAS summary data sometimes only has the magnitude of the effects, and not their sign. In this case we have two choices. We may assume that the directions of effects are the same, and do a one-tailed test of the current effect in the prior direction. Alternatively, we can do a two-tailed test by including the 2 tests (η_i, σ_i^2) and $(-\eta_i, \sigma_i^2)$ for each i, for a total of $2J$ tests. **(2)** When the prior and current trait are not both quantitative, and one is binary, the model $\nu_i = \nu_{0i}$ should be re-examined. It is still convenient as a first approximation. **(3)** Benjamini-Hochberg may be used instead of Bonferroni, with any weights summing to J, for increased power (Genovese et al., 2006).

6 Data Analysis

6.1 Data Sources

We illustrate our method by analyzing 5 publicly available genome-wide association studies of quantitative or binary traits. We use the association p-values, which are available for 500,000 to 2.5 million SNPs. The five studies are: CARDIoGRAM and C4D for coronary artery disease (Schunkert et al., 2011; Coronary Artery Disease (C4D) Genetics Consortium, 2011), and one each for the kidney trait estimated glomerular filtration rate (eGFR) creatinine (Köttgen et al., 2010), blood lipids (Teslovich et al., 2010), and schizophrenia (Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011). CARDIoGRAM and C4D include non-overlapping samples. A detailed description appears in the Supplementary material.

6.2 Specific Pairs

We analyze three pairs of data sets, with specific motivation for each. First, we use CARDIoGRAM as prior information for C4D. This is a 'positive control' for our method, since both studies measure the same phenotype, coronary artery disease. Therefore the weights should increase power. We choose C4D as target because it has smaller sample size; hence prior information may increase power more substantially.

Second, we use the blood lipids GWAS as prior information for schizophrenia. Andreassen et al. (2013) showed improved power with this pair. They developed and controlled the Bayesian conditional false discovery rate. This is not known to control a frequentist criterion. Our goal was to evaluate the power improvement using a frequentist method. As Andreassen et al. (2013) noted, there is a small overlap between the controls of the two GWAS (Section 10).

Third, we used the eGFR creatinine GWAS as prior information for the C4D coronary artery disease study. Heart disease and renal disease are comorbid (eg. Silverberg et al., 2004), so it is possible that this may improve power. Here the hypothesized improvement is not based on entirely rigorous arguments.

6.3 Methods compared

We run weighted Bonferroni multiple testing for each of the 5 weighting schemes. The prior data is $T_{0i} = \Phi^{-1}(P_{0i}/2)$, where P_{0i} is the i-th prior p-value. The family-wise error rate is controlled at $q = 0.05$.

The first four methods are: unweighted Bonferroni, where all weights equal 1; Spjotvoll weights with parameters $\mu_i = (N_i/N_{0i})^{1/2}T_{0i}$; Bayes weights defined in Section 5, with dispersion $\phi = 0.1, 1,$ and 10; and exponential weighting (Roeder et al., 2006) with tilt $\beta = 1, 2,$ and 4, introduced in Section 4.1.

The fifth and last method is filtering, which selects the smallest p-values in the prior study, and tests their SNPs in the current study. We use three p-value thresholds $P < 10^{-2}, 10^{-4}, 10^{-6}$. Rietveld et al. (2014) propose a method to choose the optimal p-value threshold for filtering. This needs the genotypic correlation between the two traits and the additive heritability of the current trait. For complex traits these parameters are usually estimated with a large uncertainty. Substantial domain expertise is required to choose the right parameter.

6.4 Additional details

We prune the significant SNPs for linkage disequilibrium (LD) using the DistiLD database (Palleja et al., 2012). We LD prune the significant SNPs for each method by selecting one SNP from each LD block. Our data analysis pipeline is available from the first author.

We compute a score $s_{m(p)d}$ for each method m with parameters p, on each data set d. This is defined as +1 if the method increases the number of hits compared to unweighted, 0 if it leaves it unchanged, and -1 otherwise. The score $s_{m(p)}$ of a method m with parameters p is the sum of scores across data sets. The total s_m of the method m is the sum of scores $s_{m(p)}$ across parameters.

6.5 Results

Table 1: Number of significant SNPs of five methods on three examples. Top: Results pruned for linkage disequilibrium (LD). Middle: results without LD pruning. Bottom: The score of each method. The methods compared are unweighted (Un); Spjotvoll (Spjot); Bayes with $\phi = 0{\cdot}1, 1, 10$; exponential (Exp) with $\beta = 1, 2, 4$ and filtering (Filter) with $-\log(P) = 2, 4, 6$. CG stands for CARDIoGRAM

	Un	Spjot	Bayes(ϕ)			Exp(β)			Filter($-\log P$)		
Parameter			$0 \cdot 1$	1	10	1	2	4	2	4	6
LD Pruned											
CG → C4D	4	11	10	8	4	4	5	4	10	10	6
Lipids → SCZ	4	1	1	1	5	1	0	0	2	2	2
eGFRcrea → C4D	4	2	2	4	4	4	5	4	1	0	1
Unpruned											
CG→ C4D	29	45	44	39	29	32	34	27	40	48	34
Lipids → SCZ	116	214	214	223	123	92	0	0	217	96	39
eGFRcrea → C4D	29	18	18	23	29	29	28	19	1	0	1
Scoring											
Score	0	0	0	1	1	0	0	−1	0	−1	−1
Total	0	0		sum = 2			sum = −1			sum = −2	

The results of our data analysis are presented in Table 1. This table has the number of significant SNPs on the pairs of GWAS data in 6.2 using the weights in 6.3. We also present the LD pruned results, which are a proxy for the the number of independent loci found.

The results are somewhat inconclusive. On the positive control example, all methods except exponential weighting improve power. Spjotvoll weighting and filtering have the largest number of SNPs. On the blood lipids example, methods generally lose power for LD pruned SNPs (except Bayes weights with $\phi = 10$); and methods can both lose and gain power for non-LD pruned SNPs (except Bayes weights which uniformly improves power). On the other hand, for the eGFR creatinine example, exponential weights show the best behavior. We also see that the default $\phi = 1$ is never worse that both unweighted and Spjotvoll, and for the unpruned lipids example it is better.

If we allow for tuning of parameters, Bayes weights show a good performance. They are either first or second in all examples, and other methods rank lower. However, since we don't have a

principled way to tune the parameters, we do not pursue this way of evaluation.

Instead, we look at the scoring method for evaluation. The only method with a positive score is our Bayes weights ($\phi = 1$ and 10). The total score, summed across parameter settings, is also only positive for Bayes testing. This shows some promise for our method. However, from this analysis alone we cannot establish conclusively the relative merits of the methods. In future work it will be necessary to evaluate p-value weighting methods on more data sets.

7 Acknowledgements

This work was supported by grants from the AFAR/EMF and the NIH/NIA (AG025941), and NSF grant DMS-1407397. The first author thanks David Donoho for support.

Supplementary material

The supplementary material is organized as follows: Accessing the software implementation of our method is described in Section 8. In Section 9 we give the proofs of all mathematical claims from the paper. In addition, in Subsection 9.5.2 we give Algorithm 2 to compute our weights. In Section 10 describe our data sources in detail.

8 Software

8.1 End-user software

We provide R and MATLAB implementations of the methods developed in this paper. They can be obtained from public git repositories or directly from the first author. The Matlab implementation is available from https://github.com/edgardobriban/pvalue_weighting_matlab. The R implementation is available from https://github.com/edgardobriban/pvalue_weighting_r. An R package is under development.

8.2 Reproducibility

All computational results and analyses of this paper have been performed in a reproducible way. To reproduce the simulation results and figures, we provide the source code in the above-mentioned Matlab package at https://github.com/edgardobriban/pvalue_weighting_matlab. To reproduce the data analyses, a separate repository has been created, and is publically available at https://bitbucket.org/edgardobriban/pvalue-weighting-gwas.

9 Proofs

9.1 Proof of Proposition 1 - Sparse Means

Proof. The objective function we need to maximize in w_i is

$$\sum_{i=1}^{J} \Phi\{\Phi^{-1}(qw_i) - \mu_i\}.$$

We know that out of the J means μ_i, $\pi_0 J$ are equal to 0. For each of these, the summand in the objective simplifies to qw_i. For the remaining $\pi_1 J$ means, the same convex objective $g(w) = \Phi\{\Phi^{-1}(qw_i) - \mu_i\}$ is summed, and the objective becomes:

$$q \sum_{i:\mu_i=0} w_i + \sum_{i:\mu_i=M} g(w_i).$$

Now, since g is convex, we have $\sum_{i:\mu_i=M} g(w_i) \leq Mg(\bar{w})$, where $\bar{w}$ is the mean of the weights of large effects $\bar{w} = (J\pi_1)^{-1}(\sum_{i:\mu_i=M} w_i)$. Hence, there is a set of optimal weights that take equal values for equal means. This proves the first claim in the proposition.

If we call w_0, w_1 the weights for 0 and M, the optimization problem takes the form

$$\max_{w_0,w_1\in[0,1/q]} \pi_0 w_0 + \pi_1 \Phi\left(\Phi^{-1}(qw_i) - M\right) \text{ s.t. } \pi_0 w_0 + \pi_1 w_1 = 1.$$

Next, recall that $qw_i \leq 1$ and introduce a new set of variables $c_i := \Phi^{-1}(qw_i)$. If $qw_i = 1$, then c_i will take the value $+\infty$ in the extended real number system $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. All our calculations respect the rules of the extended number system, so this will not cause any problems. For instance, $\Phi(\infty)$ is defined as 1 by continuity.

Then the optimization problem becomes

$$\max_{c_0,c_1\in\overline{\mathbb{R}}} \pi_0 \Phi(c_0) + \pi_1 \Phi(c_1 - M) \text{ s.t. } \pi_0 \Phi(c_0) + \pi_1 \Phi(c_1) = q.$$

Substituting the second equation we find that we need to maximize

$$q + \pi_1 (\Phi(c_1 - M) - \Phi(c_1))$$

subject to $\pi_1 \Phi(c_1) \leq q$. Now it is easy calculus to check that the function $c \to \Phi(c - M) - \Phi(c)$ is strictly increasing on $(-\infty, M/2]$ and strictly decreasing on $[M/2, \infty)$. Therefore, if $M/2$ is feasible, i.e. $\pi_1 \Phi(M/2) \leq q$, then the maximum is achieved at $c_1 = M/2$. This leads to the claimed formula for $w_1 = \Phi(c_1)/q$ and the power p^*.

If $M/2$ is not feasible, then the maximum is achieved at the largest feasible value c_1, defined by $\pi_1 \Phi(c_1) = q$. This again leads to the claimed formulas. □

9.2 Proof of Theorem 1 - Optimal weights for small q

Similarly to Section 9.1, we introduce the new set of variables $c_i := \Phi^{-1}(qw_i)$. We can equivalently rewrite our problem as

$$\max_{c\in\overline{\mathbb{R}}^J} \sum_{i=1}^{J} \Phi\left(\frac{c_i - \eta_i}{\gamma_i}\right) \text{s.t. } \sum_{i=1}^{J} \Phi(c_i) = qJ.$$

Clearly it is enough to find a scalar dual variable λ such that we can solve the Lagrangian problem

$$\max_{c\in\overline{\mathbb{R}}^J} \sum_{i=1}^{J} \Phi\left(\frac{c_i - \eta_i}{\gamma_i}\right) - \lambda\left(\sum_{i=1}^{J} \Phi(c_i) - qJ\right) \text{s.t. } \sum_{i=1}^{J} \Phi(c_i) = qJ.$$

The strategy is to study this penalized objective for each fixed λ, find maximizers $c_i(\lambda)$, and then find a suitable λ to make the constraint hold.

For this we introduce a function $f(c; \eta, \gamma)$ that is the generic term in the Lagrangian function of the problem:

$$f(c;\eta,\gamma) = \Phi\left(\frac{c-\eta}{\gamma}\right) - \lambda\Phi(c).$$

We find the maximizer of f in the following lemma.

Lemma 9.1. *If $\lambda \geq 1$, then the maximum of f is reached at*

$$c_1(\eta,\gamma;\lambda) = -\frac{\eta + \gamma\sqrt{\eta^2 + 2\left(\gamma^2-1\right)\log(\gamma\lambda)}}{\gamma^2-1}. \tag{6}$$

This function was called c in the main paper. Here we call it c_1 to distinguish it from the dummy variable c.

Proof. Denoting the standard normal density by φ, the derivative of f with respect to c is:

$$\frac{\partial f(c;\eta,\gamma)}{\partial c} = \frac{1}{\gamma}\varphi\left(\frac{c-\eta}{\gamma}\right) - \lambda\varphi(c).$$

Thus $\partial f/\partial c \geq 0$ if and only if

$$c^2 - \left(\frac{c-\eta}{\gamma}\right)^2 \geq 2\log(\gamma\lambda). \tag{7}$$

By assumption $\gamma > 1$. We have a quadratic inequality for c, and the associated quadratic equation has two roots c_1, c_2. If $\lambda \geq 1$ both roots are real: $c_1 \leq c_2$. Indeed it is easy to check that the smaller root is given in (6). Then f is increasing on $(-\infty, c_1]$ and $[c_2, \infty)$, and decreasing on $[c_1, c_2]$. Further, the limit of f at $-\infty$ is clearly 0, while at ∞ it is $1-\lambda$. Therefore, if $\lambda \geq 1$, c_1 is a global maximum of f. □

Therefore, if $\lambda \geq 1$, the i-th summand in the Lagrangian is maximized at $c_1(\eta_i, \gamma_i, \lambda)$. To solve our problem it is enough to find a $\lambda \geq 1$ such that $\sum_{i=1}^{J} \Phi(c_1(\eta_i, \gamma_i, \lambda)) = Jq$.

Such a λ exists by condition (5), as follows: c_1 is a decreasing function of λ, thus the constraint $\lambda \geq 1$ is equivalent to the sum of $\Phi(c_1)$ evaluated at $\lambda = 1$ to be greater than Jq. This is exactly the condition we assumed in (5). This finishes the proof.

9.3 Proof of Proposition 2 - Recovering the Spjotvoll weights

Let $R^2 := \left(\gamma^2-1\right)\left(\eta^2 + 2\gamma^2\log(\gamma\lambda)\right)$. Then $c_1(\eta,\gamma;\lambda)$ equals after some calculation:

$$c_1 = -\frac{\eta + \sqrt{\eta^2+R^2}}{\gamma^2-1} = -\frac{R^2}{\sqrt{\eta^2+R^2}-\eta}\frac{1}{\gamma^2-1} = -\frac{\eta^2 + 2\gamma^2\log(\gamma\lambda)}{\sqrt{\eta^2+R^2}-\eta}.$$

As $\sigma \to 0$, we have $\gamma \to 1$, $\gamma^2\log(\gamma\lambda) \to \log(\lambda)$, and $R^2 \to 0$. For $\eta < 0$ this shows that the limit of c_1 is

$$-\frac{\eta^2 + 2\log(\lambda)}{|\eta| - \eta} = \frac{\eta^2 + 2\log(\lambda)}{2\eta}.$$

Recall that the Spjotvoll weight is

$$w(\mu) = \Phi\left(\frac{\mu}{2} + \frac{c}{\mu}\right)/q.$$

This shows that the we recover the Spjotvoll weights in the limit as $\sigma \to 0$.

9.4 Proof of Proposition 3 - Simple sufficient condition

We start by showing that for $\lambda = 1$

$$c_1(\eta, \gamma; \lambda) \geq -\frac{\eta^2 + \gamma^2 \log(\gamma^2)}{2|\eta|}.$$

Indeed, for negative $\eta < 0$ the left hand side equals:

$$-\frac{\eta + \gamma\sqrt{\eta^2 + (\gamma^2 - 1)\log(\gamma^2)}}{\gamma^2 - 1} = -\frac{\eta^2 + \gamma^2 \log(\gamma^2)}{|\eta| + \gamma\sqrt{\eta^2 + (\gamma^2 - 1)\log(\gamma^2)}}$$

and now the inequality follows immediately, because $\gamma \geq 1$, so

$$\gamma\sqrt{\eta^2 + (\gamma^2 - 1)\log(\gamma^2)} \geq \gamma\sqrt{\eta^2} \geq |\eta|.$$

Thus the condition (5) of the theorem is satisfied if the following more explicit inequality holds.

$$\sum_{i=1}^{J} \Phi\left(-\frac{\eta_i^2 + \gamma_i^2 \log(\gamma_i^2)}{2|\eta_i|}\right) \geq Jq. \tag{8}$$

For this to hold it is clearly sufficient that there are M distinct indices such that (recall $\alpha = Jq$)

$$\Phi\left(-\frac{\eta_i^2 + \gamma_i^2 \log(\gamma_i^2)}{2|\eta_i|}\right) \geq \alpha/M. \tag{9}$$

The above inequality is equivalent to

$$\frac{|\eta_i|}{2} + \frac{\gamma_i^2 \log(\gamma_i^2)}{2|\eta_i|} \leq |z_{\alpha/M}|.$$

By assumption there are M indices such that $|\eta_i|/2 \leq |z_{\alpha/M}|/2$ and $\gamma_i^2 \log(\gamma_i^2)/(2|\eta_i|) \leq |z_{\alpha/M}|/2$. For these indices (9) is true, so that (8) holds. This shows that the original condition (5) holds, finishing the proof.

9.5 Proof of Theorem 2 - Optimal weights in the general case

The proof builds on Theorem 1, but requires more detailed analysis. There are two parts: understanding the monotonicity of the generic term in the Lagrangian, then using it to find the optimal weights.

9.5.1 Monotonicity of the generic term

We saw in Lemma 9.1 that if $\lambda \geq 1$, f is maximized at c_1. We also saw in the proof that there are two cases: the roots of (7) are either real or complex. If the roots are complex, then f is increasing and the supremum is at $+\infty$. The two roots are complex when the discriminant of the quadratic equation is negative:

$$\eta^2 + 2\left(\gamma^2 - 1\right)\log(\gamma\lambda) < 0.$$

This inequality is equivalent to the following bound for λ, in terms of a new function $l(\eta, \gamma)$:

$$\lambda < l(\eta, \gamma) := \frac{1}{\gamma} \exp\left(-\frac{\eta^2}{2(\gamma^2 - 1)}\right). \tag{10}$$

Therefore if $\lambda < l(\eta, \gamma)$, then the supremum of f is at $c = +\infty$. Otherwise there are two candidates for the supremum: $c = +\infty$ and c_1. We want to compare the two candidate maxima.

Denote the limit of f at infinity by $p_\infty(\lambda) = p_\infty(\lambda; \eta, \gamma)$ and the value of f at c_1, when defined, as $p_1(\lambda) = p_1(\lambda; \eta, \gamma) = f(c_1)$. Let the difference between the two extrema be $d(\lambda; \eta, \gamma) = p_1(\lambda; \eta, \gamma) - p_\infty(\lambda, \eta, \gamma)$. In the following lemma we find the maximizer of f as a function of λ.

Lemma 9.2. *There is a unique value $k(\eta, \gamma)$ in the interval $[l(\eta, \gamma), 1]$ such that $d(k(\eta, \gamma); \eta, \gamma) = 0$. For $\lambda < k(\eta, \gamma)$, the supremum of f is at $+\infty$, else it is at c_1. For the special value $\lambda = k(\eta, \gamma)$ both values are equal.*

Proof. We have explicitly $p_\infty(\lambda) = 1 - \lambda$ and thus p_∞ decreases linearly from 1 to 0 on the unit interval $[0, 1]$. We also have by definition

$$p_1(\lambda) = \Phi\left(\frac{c_1(\eta, \gamma; \lambda) - \eta}{\gamma}\right) - \lambda\Phi(c_1(\eta, \gamma; \lambda)).$$

Differentiating this expression with respect to λ reveals

$$\frac{\partial p_1}{\partial \lambda} = \frac{1}{\gamma}\varphi\left(\frac{c_1 - \eta}{\gamma}\right)\frac{\partial c_1}{\partial \lambda} - \left(\lambda\varphi(c_1)\frac{\partial c_1}{\partial \lambda} + \Phi(c_1)\right).$$

However, by the definition of c_1 we have $\varphi((c_1 - \eta)/\gamma) - \lambda\varphi(c_1) = 0$. Indeed, c_1 was defined as one of the extrema of f, which leads to the above equation by taking derivatives. Hence the expression above simplifies to: $\partial p_1/\partial\lambda = -\Phi(c_1)$.

This shows that p_1 is decreasing in λ: the derivative belongs to $(-1, 0)$. We also know that $p_1 > 0$, because the value p_1 is a local maximum of f in c, and we have seen that f - as a function of c - increases from 0 (a value which it takes in the limit at $-\infty$) to $f(c_1)$ - and hence $p_1 = f(c_1)$ must necessarily be positive. We conclude that $p_1(1) > 0 = p_\infty(1)$

Finally, we note that $p_1(\lambda)$ is well-defined precisely when $c_1(\lambda)$ is. This happens when the expression inside the square root is non-negative, which means $\lambda \geq l(\eta, \gamma)$.

To summarize

- d is defined on the interval $\lambda \in [l(\eta, \gamma), 1]$.
- d is a strictly increasing differentiable function on this interval, because its derivative is $\frac{\partial d}{\partial \lambda} = 1 - \Phi(c_1) > 0$.
- $d(1) > 0$.

If we show that d has a unique root k, our conclusion will follow. Given the above three statements, it's enough to show that $d(l(\eta, \gamma)) < 0$; then the claim follows by the intermediate value theorem.

To check the condition $d(l(\eta, \gamma)) < 0$ we note that $c_1(l(\eta, \gamma)) = -\frac{\eta}{\gamma^2 - 1}$, hence $p_1(l) < p_\infty(l)$ is equivalent to

$$\Phi\left(-\frac{\eta\gamma}{\gamma^2 - 1}\right) - l\Phi\left(-\frac{\eta}{\gamma^2 - 1}\right) < 1 - l,$$

or after some rearrangement $l\Phi\left(\frac{\eta}{\gamma^2 - 1}\right) < \Phi\left(\frac{\eta\gamma}{\gamma^2 - 1}\right)$.

Introducing the variable $x = \frac{\eta}{\gamma^2 - 1}$ and eliminating η by noting $l = \exp(-x^2(\gamma^2 - 1)/2)/\gamma$ we obtain that all we need to show is

$$\exp\left(\frac{x^2}{2}\right)\Phi(x) \leq \exp\left(\frac{(\gamma x)^2}{2}\right)\Phi(\gamma x)\gamma.$$

If $\eta < 0$, then $x < 0$, and the above will follow if the function $m(x) = x\exp\left(\frac{x^2}{2}\right)\Phi(x)$ is strictly increasing on $(-\infty, 0]$. Checking this is an elementary calculus exercise. The cases $\eta = 0$ and $\eta > 0$ are handled similarly. This finishes the proof.

□

9.5.2 Method for computing weights

In the previous section, we saw that for any i, if $\lambda < k_i := k(\eta_i, \gamma_i)$, then the supremum of $f(\cdot; \eta_i, \gamma_i)$ occurs at $+\infty$, else it occurs at $c_1(\eta_i, \gamma_i; \lambda)$. For the optimal weights this shows:

$$w_i(\eta_i, \gamma_i; \lambda) = \begin{cases} \Phi(c_1(\eta_i, \gamma_i; \lambda))/q & \text{if } \lambda > k(\eta_i, \gamma_i), \\ \Phi(c_1(\eta_i, \gamma_i; \lambda))/q \text{ or } 1/q & \text{if } \lambda = k(\eta_i, \gamma_i), \\ 1/q & \text{if } \lambda < k(\eta_i, \gamma_i). \end{cases}$$

We emphasize that the value of w_i can take either of two values for $\lambda = k(\eta_i, \gamma_i)$. We will unambiguously specify a choice later for each i, determined by the constraint on the sum of the weights.

Now all that remains is to search for a suitable λ such that the weights sum to J: $\sum_i w_i = J$. If we find such a λ, then by duality the weights w_i will solve our original problem. Let us denote the sum of the weights by W:

$$W(\lambda) = \sum_{i=1}^{J} w_i(\eta_i, \gamma_i; \lambda).$$

The function $W(\lambda)$ is unambiguously defined for $\lambda \neq k_i$, which is what we will use at first.

Each function w_i is decreasing in λ: constant on the interval $(0, k_i)$, and decreasing smoothly from $r_i = \Phi(c_1(\eta_i, \gamma_i; k_i))/q$ to 0 on the interval (k_i, ∞). Further, the function w_i has a jump of size $(1 - r_i)/q$ at k_i.

Therefore W is a decreasing function of λ on $[0, \infty)$, going from J/q to 0. Further, if we sort the values k_i such that $k_{(0)} = 0 < k_{(1)} \leq \ldots \leq k_{(J)} < k_{(J+1)} = \infty$, then W is smooth on the intervals $(k_{(i)}, k_{(i+1)})$, and has jumps of size $(1 - r_{(i)})/q$ at $k_{(i)}$ for $1 \leq i \leq J$.

Hence, for our problem of solving $W(\lambda) = J$, there are two possibilities:

Case 1 There is an interval $(k_{(i)}, k_{(i+1)})$ such that J belongs to the image of this interval under W. In this case, since the function W is strictly decreasing and continuous on this interval, there will exist a unique λ in the interval such that $W(\lambda) = J$. In this case we can solve the original problem exactly.

Case 2 There is no such interval. In this case, the present duality approach is unable to produce an exact solution to the original problem.

However, we can get an approximate solution. Let us consider the values $W(k_i)$. Above we noted that $w_i(k_i)$ can take two possible values $1/q$ and r_i. Hence W can also take more than one value at k_i. Since several k_i may be equal, W may take more than two possible values, because each summand w_i can be chosen in two ways. To understand this, let us call the distinct values of k_i to be K_j, and assume $K_1 < K_2 < \ldots < K_L$. For brevity let us also add $K_0 = 0$, $K_{L+1} = \infty$. Without loss of generality, we assume that the values k_i cluster in the following way:

$K_1 = k_{(1)} = k_{(2)} = \ldots = k_{(i_1)}$, $K_2 = k_{(i_1+1)} = \ldots = k_{(i_2)}$, and so on. We also define the sets S_i containing the indices j for which $k_j = K_i$.

The value of W is now defined unambiguously on each interval (K_i, K_{i+1}), and W is a smooth decreasing function on this interval. At K_m, W has a jump of size

$$j_m = \sum_{i:k_i=K_m} \frac{1-r_{(i)}}{q} = \frac{i_{m+1} - i_m - \sum_{i=i_m+1}^{i_{m+1}} r_{(i)}}{q}.$$

By assumption, in the current case there is no interval (K_i, K_{i+1}) whose image under W contains J. Therefore J is contained in one of the jumps occurring between these intervals, say the jump at K_i. Let us now choose $\lambda = K_i$. Let us also define the left and right limits of W at K_i: $a_+ = \lim_{x \to K_i, x<K_i} W(x)$, $a_- = \lim_{x \to K_i, x>K_i} W(x)$.

Then $a_- < a_+$ and the current case entails that $J \in [a_-, a_+]$. Let us also denote the values of the jumps by the more compact notation $s_i = (1 - r_{(i)})/q$. With these notations, we can now easily state that the possible values of $W(K_i)$ are, for any set $S \subset S_i$

$$W(K_i, S) = a_- + \sum_{t \in S} s_t.$$

These consitute at most $2^{|S_i|}$ different values, but some of them may be equal. Nonetheless, the values s_i are between 0 and $1/q$, so for any value $x \in [a_-, a_+]$, we can find a suitable subset S such that $W^* = W(K_i, S)$ approximates x within $1/(2q)$. In fact, for any ordering of the s_t values $s_{t_1}, s_{t_2}, \ldots, s_{t_{|S_i|}}$, looking at the two cumulative sums $s_{t_1} + s_{t_2} + \ldots + s_{t_k}$ scoring nearest to x, we can find one closer than $1/(2q)$. This shows that the desired subset can be found in $O(|S_i|)$ steps; with more work we can possibly find a better packing.

For this choice of the values w_i, the sum of w_i equals W^*, and $|W^* - J| \leq 1/(2q)$. To summarize this second case: Even if we can't find a dual variable λ such that the constraint is satisfied, we can find one such that the sum of w_i equals W^*, and $|W^* - J| \leq 1/(2q)$.

In fact we have found the exact optimal weights for a slightly different problem. Define the scaled weights $\widetilde{w_i} = Jw_i/W^*$, then $\widetilde{w_i}$ solves the optimization problem 3 with the parameter q replaced by $q^* = W^*q/J$. This is clear if we notice that the optimization problem can be parametrized by q, and everything we've said so far holds for each fixed q. Thus, for any value of W^*, we just need to find a value q^* such that the constraint holds. This amounts to

$$\sum_{i=1}^{J} \Phi(c_i) = W^*q = q^*J$$

and we find the claimed equation for q^*. Further $|q^* - q| = |q(J - W^*)/J| \leq 1/(2J)$, as desired. This finishes the proof that q^* exists, we just need to show how to find it.

Final algorithm The final algorithm is sumarized in Algorithm 2. We compute the values k_i for each i, using Brent's method on the equation $d(k) = 0$. This takes $O(J)$ steps. Next, we sort the unique values k_i ($O(J \log J)$ steps), and do a binary search on the values $W(k_i)$ to find the right interval for λ. Calculating a value $W(\lambda)$ takes $O(J)$ steps. The binary search takes $O(\log J)$ function evaluations of W, making the total cost of this step $O(J \log J)$.

Once we found the right interval, we have to deal with the two possibilities identified above. If there is an exact solution $W(\lambda) = J$, then after finding the right interval, Brent's method is used to solve the equation ($O(J)$ steps). On this interval, the function W is smooth with explicitly computable derivatives. If there is no exact solution, then we simply return the closest interval endpoint ($O(J)$ steps). The overall cost is $O(J \log J)$.

Algorithm 2 Method to compute weights

$\eta_i, \sigma_i^2 \leftarrow$ prior means and variances $i = 1, \ldots, J$
$q \leftarrow$ significance threshold
if condition (5) holds **then**
    Solve $\sum_i \Phi(c_1(\eta_i, \gamma_i, \lambda)) = Jq$ for $\lambda \in [1, \infty)$ using Newton's method
    return $w_i = \Phi(c_1(\eta_i, \gamma_i, \lambda))/q$, $q^* = q$
else
    Define $d(\lambda; \eta, \gamma) = \lambda\Phi(-c_1(\eta, \gamma; \lambda)) - \Phi\left(-(c_1(\eta, \gamma; \lambda) - \eta)/\gamma\right)$
    for all i Solve $d(k_i; \eta_i, \gamma_i) = 0$ for $k_i \in [l(\eta_i, \gamma_i), 1]$ using Brent
    Let the sorted unique values k_i be $K_{(i)}$
    Define $w_i^-(\lambda) = \begin{cases} \Phi(c_1(\eta_i, \gamma_i; \lambda))/q & \text{if } \lambda \geq k_i \\ 1/q & \text{if } \lambda < k_i \end{cases}$
    Define $s_i(\lambda) = I(\lambda = k_i)(1/q - w_i^-(\lambda))$
    Define $w_i^+(\lambda) = w_i^-(\lambda) + s_i(\lambda)$
    Define $W^+(\lambda) = \sum_i w_i^+(\lambda)$ and $W^-(\lambda) = \sum_i w_i^-(\lambda)$
    Find the index j for which $W^+(K_{(j)}) \geq Jq > W^+(K_{(j+1)})$ via binary search
    if $W^-(K_{(j)}) \geq Jq > W^+(K_{(j+1)})$ **then**
        Solve $W^+(\lambda) = Jq$ for $\lambda \in [K_{(j)}, K_{(j+1)})$ using Brent
        return $w_i = w_i^+(\lambda)$, $q^* = q$
    else
        Find the indices $S = \{j_1, \ldots, j_n\}$ such that $k_{j_i} = K_{(i)}$
        Find the largest index T such that $r^- := W^-(K_{(j)}) + \sum_{i \leq T} S_{j_i}(K_{(j)}) \leq Jq$
        Define $r^+ = r^- + S_{j_{T+1}}(K_{(j)})$
        Let W^* be the closer of r^-, r^+ to Jq (break ties arbitrarily)
        Define $w_i^* = \begin{cases} w_i^+(K_{(j)}) & \text{if } i \notin S \\ w_i^+(K_{(j)}) & \text{if } i = j_k, \text{ for } k > T+1 \\ w_i^-(K_{(j)}) & \text{if } i = j_k, \text{ for } k \leq T \\ w_i^\pm(K_{(j)}) & \text{if } i = j_{T+1}, \text{ depending on the choice of } r^\pm \end{cases}$
        return $\widetilde{w_i} = Jw_i^*/W^*$, $q^* = W^*q/J$.

10 Data sources

10.1 CARDIoGRAM - CAD

This is a meta-analysis of 14 coronary artery disease (CAD) GWAS, comprising 22,233 cases and 64,762 controls of European descent (Schunkert et al., 2011). The study includes 2.3 million SNPs. In each of the 14 studies and for each SNP, a logistic regression of CAD status was performed on the number of copies of one allele, along with suitable controlling covariates. The resulting effect sizes were combined across studies using fixed effects or random effects meta-analysis with inverse variance weighting.

10.2 C4D - CAD

This is a meta-analysis of 5 heart disease GWAS, totalling 15,420 CAD cases and and 15,062 controls (Coronary Artery Disease (C4D) Genetics Consortium, 2011). The samples did not overlap those from CARDIoGRAM. The analysis steps were similar to CARDIoGRAM.

10.3 Chronic Kidney Disease Consortium (CKDGen) - eGFR creatinine

This is a GWAS of kidney traits in 67,093 participants of European ancestry from 20 population-based cohorts (Köttgen et al., 2010). eGFR creatinine (eGFRcrea) was the trait with the largest sample size. There is no reported overlap with the samples from C4D. The analysis steps were similar to the previous two studies.

10.4 Blood Lipids

This is a GWAS of blood lipids in a sample from European populations (Teslovich et al., 2010). Triglyceride levels (TG) were one of the traits, with sample size 96,598, chosen here out of all lipids because of its previous appearance in (Andreassen et al., 2013). Standard protocols for GWAS were used: linear regression analysis with study-specific covariates, combined using fixed-effects meta-analysis.

10.5 Psychiatric Genomics Consortium - Schizophrenia

This is a mega-analysis (i.e. using the raw data not just summaries) combining GWAS data from 17 separate studies of schizophrenia (SCZ), with a total of 9,394 cases and 12,462 controls (Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011). They tested for association using logistic regression of SCZ status on the allelic dosages. The overlap with the blood lipids study consists of 1,459 controls (12% of controls in the SCZ study), from the British 1958 Birth Cohort of the Wellcome Trust Case Control Consortium.

References

Andreassen, O. A., Djurovic, S., Thompson, W. K., Schork, A. J., Kendler, K. S., ODonovan, M. C., Rujescu, D., Werge, T., van de Bunt, M., Morris, A. P., et al. (2013). Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *The American Journal of Human Genetics*, 92(2):197–209.

Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418.

Bourgon, R., Gentleman, R., and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107(21):9546–9551.

Box, G. E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, pages 383–430.

Brooks-Wilson, A. R. (2013). Genetics of healthy aging and longevity. *Human genetics*, 132(12):1323–1338.

Coronary Artery Disease (C4D) Genetics Consortium (2011). A genome-wide association study in europeans and south asians identifies five new loci for coronary artery disease. *Nature genetics*, 43(4):339–344.

Darnell, G., Duong, D., Han, B., and Eskin, E. (2012). Incorporating prior information into association studies. *Bioinformatics*, 28(12):i147–i153.

Eskin, E. (2008). Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome research*, 18(4):653–660.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, 93(3):509–524.

Gui, J., Tosteson, T. D., and Borsuk, M. E. (2012). Weighted multiple testing procedures for genomic studies. *BioData mining*, 5(4).

Hjelmborg, J., Iachine, I., Skytthe, A., Vaupel, J. W., McGue, M., Koskenvuo, M., Kaprio, J., Pedersen, N. L., Christensen, K., et al. (2006). Genetic influence on human lifespan and longevity. *Human genetics*, 119(3):312–321.

Köttgen, A., Pattaro, C., Böger, C. A., Fuchsberger, C., Olden, M., Glazer, N. L., Parsa, A., Gao, X., Yang, Q., Smith, A. V., et al. (2010). New loci associated with kidney function and chronic kidney disease. *Nature genetics*, 42(5):376–384.

Lynch, M. and Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sinauer Sunderland.

Palleja, A., Horn, H., Eliasson, S., and Jensen, L. J. (2012). Distild database: diseases and traits in linkage disequilibrium blocks. *Nucleic acids research*, 40(D1):D1036–D1040.

Rietveld, C. A., Esko, T., Davies, G., Pers, T. H., Turley, P., Benyamin, B., Chabris, C. F., Emilsson, V., Johnson, A. D., Lee, J. J., et al. (2014). Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proceedings of the National Academy of Sciences*, 111(38):13790–13794.

Roeder, K., Bacanu, S.-A., Wasserman, L., and Devlin, B. (2006). Using linkage genome scans to improve power of association in genome scans. *The American Journal of Human Genetics*, 78(2):243–252.

Roeder, K. and Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 24(4):398–413.

Rubin, D., Dudoit, S., and Van der Laan, M. (2006). A method to increase the power of multiple testing procedures through sample splitting. *Statistical Applications in Genetics and Molecular Biology*, 5(1):1–19.

Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature genetics*, 43(10):969–976.

Schunkert, H., König, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., Preuss, M., Stewart, A. F., Barbalic, M., Gieger, C., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics*, 43(4):333–338.

Silverberg, D., Wexler, D., Blum, M., Schwartz, D., and Iaina, A. (2004). The association between congestive heart failure and chronic renal disease. *Current opinion in nephrology and hypertension*, 13(2):163–170.

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495.

Spjotvoll, E. (1972). On the optimality of some multiple comparison procedures. *The Annals of Mathematical Statistics*, 43(2):398–411.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713.

Wasserman, L. and Roeder, K. (2006). Weighted hypothesis testing. *arXiv preprint math/0604172*.

Westfall, P. H., Krishen, A., and Young, S. S. (1998). Using prior information to allocate significance levels for multiple endpoints. *Statistics in medicine*, 17(18):2107–2119.