Data-driven hypothesis weighting increases detection power in genome-scale multiple testing

Nikolaos Ignatiadis, Bernd Klaus, Judith B Zaugg & Wolfgang Huber

Hypothesis weighting improves the power of large-scale multiple testing. We describe independent hypothesis weighting (IHW), a method that assigns weights using covariates independent of the *P*-values under the null hypothesis but informative of each test's power or prior probability of the null hypothesis (http://www.bioconductor.org/packages/IHW). IHW increases power while controlling the false discovery rate and is a practical approach to discovering associations in genomics, high-throughput biology and other large data sets.

Multiple hypothesis testing is an important part of many high-throughput data analysis workflows. A common objective is to maximize the number of discoveries while controlling the expected fraction of false discoveries, known as the false discovery rate (FDR). Commonly used procedures, such as that of Benjamini and Hochberg (BH)[1], achieve this objective by working solely off the list of *P*-values for individual tests[1–5]. However, these approaches have suboptimal power when the individual tests differ in statistical properties such as sample size, true effect size, signal-to-noise ratio or prior probability of being false.

For example, in differential expression analysis of RNA-seq data, tests are performed on individual genes, which can differ greatly in the number of mapped reads and the corresponding signal-to-noise ratio. In genome-wide association studies (GWAS), the power to detect associations between genetic variants and traits is lower for rare polymorphisms (all else being equal). In expression quantitative trait loci (eQTL) mapping, *cis* effects are *a priori* more likely than associations between a gene product and a distant polymorphism.

To take into account such differences in the statistical properties of the tests, one can associate each test with a weight (**Supplementary Note 1**). Weights are non-negative and fulfill a budget criterion, commonly that they average to 1; hypotheses with higher weights are prioritized[6]. The procedure of Benjamini and Hochberg[1] can be modified to allow weighting simply by replacing the original *P*-values P_i with their weighted versions P_i/w_i (where w_i is the weight of hypothesis i)[6]. This approach is known to control the FDR if the weights are prespecified and thus independent of the data. However, the optimal choice of weights is rarely known in practice, and a generally applicable data-driven method is desirable[7–11].

We developed IHW as a multiple testing procedure that applies the weighted BH method[6] using weights derived from the data (Online Methods and **Supplementary Note 2**). The input to IHW is a two-column table of *P*-values and covariates. The covariate can be any continuous or categorical variable that is thought to provide information on the statistical properties of the hypothesis tests while remaining independent of the *P*-value under the null hypothesis[9].

Such covariates exist in many applications and are often apparent to domain experts (**Table 1**). The conditional independence property can be verified either mathematically[9] or empirically[12]. Simple diagnostic plots of the data can help assess these assumptions. For example, a histogram of all *P*-values will typically show a mixture of a uniform distribution (corresponding to the true null hypotheses) and an enrichment of small *P*-values to the left (corresponding to the alternatives) (**Fig. 1a**). Splitting the hypotheses into groups based on the values of a good covariate will alter the proportion and/or the shape of the alternative distribution between the groups (**Fig. 1b–d**). If all histograms look the same, the covariate is uninformative, and its use will not lead to an increase in power. If the tails are no longer uniform, independence under the null is violated, and application of IHW is not valid.

IHW is motivated by considering multiple testing as a resource allocation problem[6]: given a budget of acceptable FDR, how can it be distributed among the hypotheses in such a way as to obtain the best possible power overall? The first idea is to use the covariate to assign hypothesis weights. We approximate the covariate–weight relationship by a stepwise constant function. No further assumptions (e.g., monotonicity) are needed. The second idea is that the number of discoveries of the weighted BH procedure with given weights is an empirical indicator of the method's power. Therefore, a good choice of the covariate–weight function should lead to a high number of discoveries.

The basic steps of IHW are as follows. First, we divide the tests into groups based on the covariate. Each group is associated with a weight so that all hypotheses within a group are assigned the same weight. For each possible choice of weights, we apply the weighted BH procedure at level α and calculate the total number of discoveries. We choose the weights leading to the highest number of discoveries.

For many applications, this approach ('naive IHW') provides satisfactory results but it has two shortcomings: first, the underlying optimization problem is difficult and does not easily scale to problems with millions of tests. Second, the naive IHW approach

European Molecular Biology Laboratory, Heidelberg, Germany. Correspondence should be addressed to W.H. (whuber@embl.de).

RECEIVED 2 DECEMBER 2015; ACCEPTED 3 MAY 2016; PUBLISHED ONLINE 30 MAY 2016; DOI:10.1038/NMETH.3885

Table 1 | Examples of covariates

Application	Covariate
Differential expression	Sum of read counts per gene across all samples[12]
GWAS	Minor allele frequency
eQTL, chromatin immunoprecipitation–QTL	Distance between genetic variant and locus of expression, or comembership in a topologically associated domain[16]
t-test	Overall variance[9]
Two-sided tests	Sign of the effect
Various applications	Signal quality, sample size

leads to loss of type I error control in certain situations for reasons analogous to overfitting in statistical learning. We use methods from statistical learning—convex relaxation, data splitting and regularization—to overcome these shortcomings in the full IHW algorithm (Online Methods and **Supplementary Note 2**).

IHW has a greater empirical detection power than the BH procedure, as we illustrate for three exemplary applications (**Supplementary Note 3**). The first is an RNA-seq data set used to detect differential gene expression between mouse strains[13,14] based on *P*-values calculated with DESeq2 (ref. 12). Here we used the mean of normalized counts for each gene, across samples, as the informative covariate, and we saw an increased number of discoveries compared with those of BH (**Fig. 2a**). The learned weight function prioritized genes with higher mean normalized counts (**Supplementary Fig. 1a**).

Second, we analyzed a quantitative mass-spectrometry (hyperplexed) experiment in which yeast cells treated with rapamycin were compared to yeast cells treated with dimethyl sulfoxide (2×6 biological replicates)[15]. Differential abundance of 2,666 proteins was evaluated using Welch's *t*-test[15]. As a covariate, we used the total number of peptides that was quantified across all samples for each protein. IHW again showed increased power compared with that of BH (**Fig. 2b**), and proteins with more quantified peptides were assigned higher weight, as expected (**Supplementary Fig. 1b**).

In a third example, we searched for associations between SNPs and histone modifications (H3K27ac) (ref. 16) on human chromosome 21. This yielded 180 million tests. As a covariate, we used the genomic distance between the SNP and the ChIP-seq signal. The power increase compared with that of BH was dramatic (**Fig. 2c**). IHW automatically assigned most weight to small distances (**Fig. 2d**). Thus, IHW acted similarly to the common practice in eQTL analysis of searching for associations only within a certain distance, a form of independent filtering. However, IHW had the advantage that no arbitrary choice of distance threshold was needed, and the weights were more nuanced than a hard distance threshold. IHW does not exclude distant SNP–phenotype pairs, which can still be detected given a sufficiently small *P*-value.

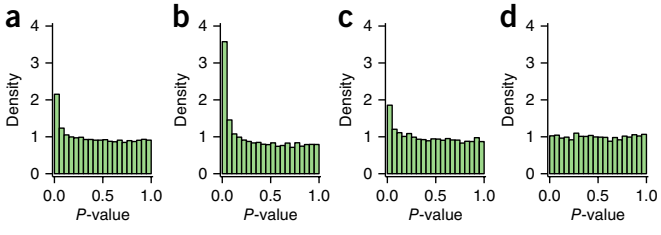


Figure 1 | Histograms stratified by the covariate as a diagnostic plot. (**a**) The histogram of all *P*-values shows a mixture of a uniform distribution and an enrichment of small *P*-values to the left. Such a well-calibrated histogram is the starting point for most multiple testing methods. (**b**–**d**) Histograms after splitting the hypotheses into three groups based on the values of the covariate.

Naive IHW, as well as previous approaches to covariate adjusted multiple testing, do not maintain FDR control in situations where all hypotheses are true (**Fig. 2e**) or where there is insufficient power to detect the false hypotheses (**Supplementary Fig. 2a**). In addition, local true discovery rate (tdr) methods (Clfdr and FDRreg) often show strong deviations from the target FDR in a direction (conservative or anticonservative) that is not apparent *a priori* (**Fig. 2f,g**). Thus, among all methods benchmarked across these scenarios, only BH, IHW (but not naive IHW) and LSL-GBH generally control the FDR (**Fig. 2** and **Supplementary Fig. 2**; summarized in **Supplementary Table 1**; simulations described in **Supplementary Note 4**).

IHW can apply a size investing strategy. IHW already assigns low weight to covariate groups with low signal (such as in **Fig. 1d**), but a less intuitive effect can pertain to groups with very small *P*-values. Size investing[17] is a strategy by which IHW can shift weight from these groups with small *P*-values toward groups with more intermediate *P*-values, since the former will be rejected even with a lower weight. Several other methods (**Supplementary Table 1**), including greedy independent filtering, stratified BH, LSL-GBH, TST-GBH and FDRreg, cannot apply size investing and can even lose power compared with the BH method in situations where size investing would be beneficial (**Supplementary Fig. 2d,f** and **Supplementary Note 5**).

It is instructive to consider the relation between IHW and the concept of tdr. *P*-values are a reduction of data into one number, which typically does not contain all the important information (**Table 1**; refs. 18 and 19). One might wonder whether there are other quantities that are better suited for selecting discoveries. The theoretically optimal candidate is the tdr (ref. 4), defined for the *i*th hypothesis as

$$\mathrm{tdr}_i(p) = \pi_{1,i}\frac{f_{1,i}(p)}{f_i(p)} \qquad (1)$$

where f_i is the density of the distribution of the *P*-value P_i (see **Fig. 3a** for explanation, as well as **Supplementary Figs. 3** and **4**). f_i is a mixture of two densities, $f_i = \pi_{0,i}f_0 + \pi_{1,i}f_{1,i}$, where f_0 and $f_{1,i}$ are conditional on the null or the alternative hypothesis being true, respectively, and $\pi_{0,i}$ and $\pi_{1,i}$ (which sum up to 1) are the corresponding prior probabilities. The null distribution of a properly calibrated test is uniform, therefore we can set $f_0(p) = 1$ irrespective of p and i. We give examples of three hypotheses with different tdr curves (**Fig. 3b–d**).

It can now be shown that, to maximize power at a given FDR, one should reject the hypotheses with the highest tdr (refs. 20 and 21). In other words, if we knew the functions in equation (1) and could use $\mathrm{tdr}_i(P_i)$ as our test statistics, then without any further effort we would have a method for FDR control with optimal power.

Similarly to the central idea of IHW, one might assume that the many different, unknown univariate functions $\mathrm{tdr}_i(p)$, one for each hypothesis i, can be approximated by a single bivariate function $\mathrm{tdr}(p, x)$, where x is the value of the covariate X. The joint density of P and X (**Fig. 3e**) gives rise to the joint density of tdr and X (**Fig. 3f**). In such a scenario, the decision boundary of the

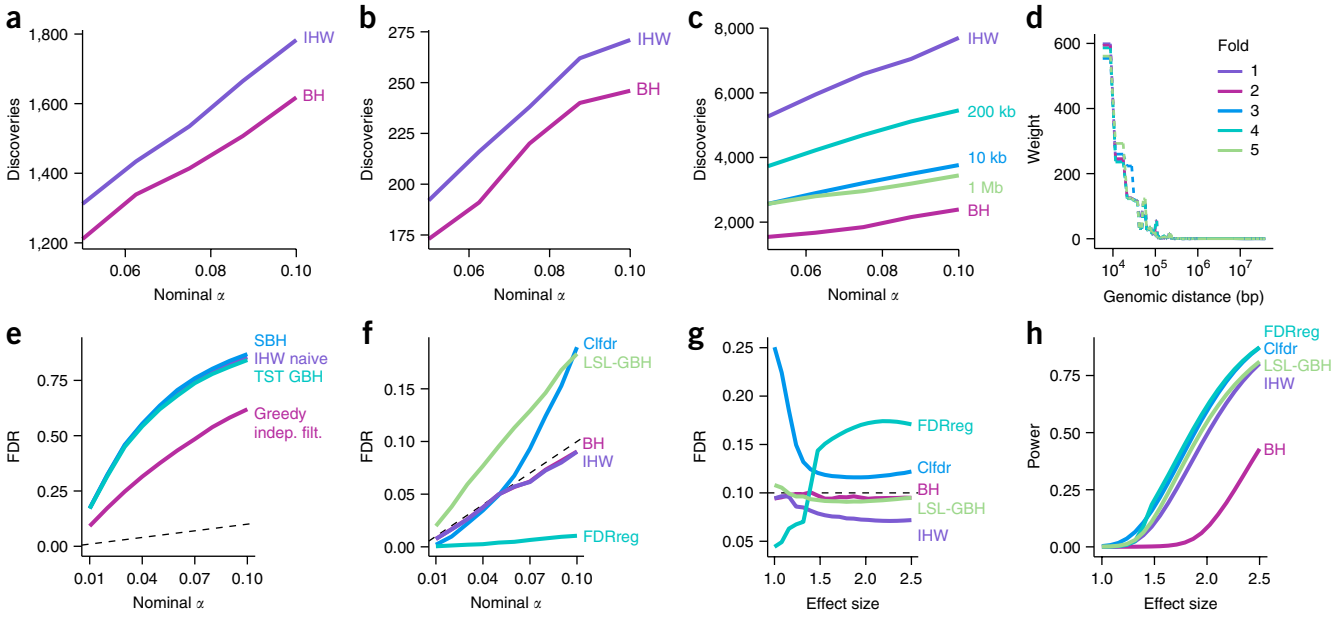
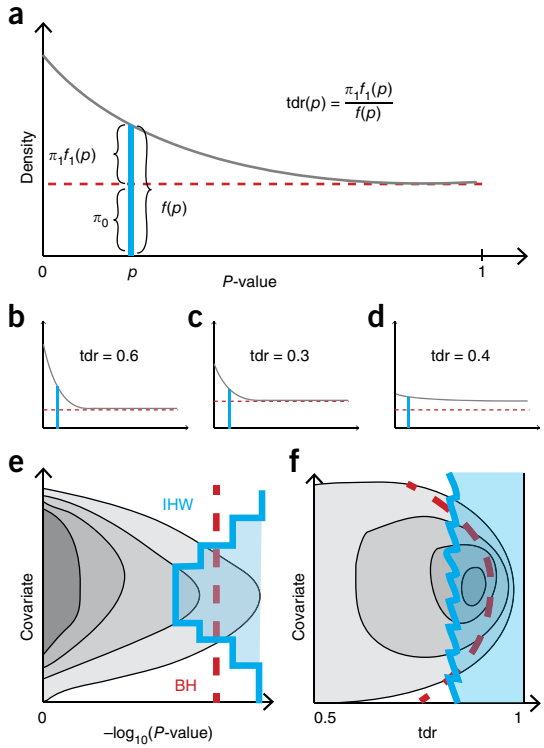


Figure 2 | Performance evaluation. (**a**–**c**) Number of discoveries as a function of the target FDR for (**a**) RNA-seq data[13] with mean of normalized counts for each gene as the covariate. (**b**,**c**) Hyperplexed mass-spectrometry data[15], with number of peptides quantified per protein as the covariate (**b**), and histone QTL (hQTL) data set[16] for chromosome 21, with genomic distance between SNPs and ChIP-seq signals as the covariate (**c**). Independent filtering with different distance cutoffs was also applied. (**d**) Weight function learned by IHW at $\alpha = 0.1$ for the hQTL data set. Curves represent the five folds in the data-splitting scheme. (**e**–**h**) Performance on simulated data (see **Supplementary Table 1** for descriptions of methods). (**e**,**f**) Type I error control if all null hypotheses are true. (**e**) All methods shown make too many false discoveries. (**f**) BH, FDRreg, and IHW control the FDR. LSL-GBH and Clfdr are slightly anticonservative. (**g**,**h**) Implications of different effect sizes. The two-sample *t*-test was applied to Normal samples ($n = 2 \times 5$, $\sigma = 1$) with either the same mean (nulls) or means differing by the effect size indicated on the *x*-axis (alternatives). The fraction of alternatives was 0.05. The pooled sample variance was used as the covariate. The nominal level was $\alpha = 0.1$ (dotted line). (**g**) The *y*-axis shows the actual FDR (dotted line refers to nominal level). (**h**) Power analysis. All methods show improvement over BH.

BH method tends to be suboptimal as it is defined solely in terms of *P*-values (**Fig. 3e**) and thus differs from the optimal region, whose boundary is a vertical line of constant tdr (**Fig. 3f**).



In practice, however, we neither know the quantities in equation (1) nor the bivariate function tdr(p, x) and have to estimate them[22]. Unfortunately, this estimation problem is difficult, and even with the use of additional approximations, such as splines[23] or piecewise constant functions[24], there does not seem to be a practical implementation.

An important feature of IHW is that it circumvents explicit estimation of the bivariate tdr function and instead yields a powerful testing procedure by working directly on *P*-values and covariates to assign data-driven hypothesis weights. In addition, the IHW method readily extends to other weighted multiple testing

Figure 3 | Local true discovery rate and informative covariates. (**a**) Schematic representation of the density f_i, which is composed of the alternative density $f_{1,i}$ weighted by its prior probability $\pi_{1,i}$ and the uniform null density weighted by $\pi_{0,i}$. (**b**–**d**) The tdr of individual tests can vary. In **b**, the test has high power, and $\pi_{0,i}$ is well below 1. In **c**, the test has equal power, but $\pi_{0,i}$ is higher, leading to a reduced tdr. In **d**, $\pi_{0,i}$ is as in **b**, but the test has little power, again leading to a reduced tdr. (**e**) If an informative covariate is associated with each test, the distribution of the *P*-values from multiple tests is different for different values of the covariate. The contours represent the joint density of *P*-values and covariate. The BH procedure accounts only for the *P*-values and not the covariates (dashed red line). In contrast, the decision boundary of IHW is a step function: each step corresponds to one group, i.e., to one weight. (**f**) According to equation (1), the density of the tdr also depends on the covariate. The decision boundary of the BH procedure (dashed red line) leads to a suboptimal set of discoveries, in this example with higher than optimal tdr for intermediate covariate values and lower than optimal tdr for other values. In contrast, IHW approximates a line of constant tdr, implying efficient use of the FDR budget.

procedures[6] including IHW–Bonferroni (**Supplementary Note 6** and **Supplementary Fig. 5**), a new, powerful method for controlling the familywise error rate (FWER). In contrast, local tdr methods are specific to the FDR.

We have introduced a weighted multiple testing method that learns the weights from the data. Its appeal lies in its generic applicability. It does not require assumptions about the relationship between the covariate and the power of the individual tests, such as monotonicity, which is necessary for independent filtering. It can apply size investing strategies, since it does not assume that the alternative distributions are the same across the different hypotheses. Furthermore, IHW is computationally robust and scales to millions of hypotheses.

The idea of using informative covariates for hypothesis weighting or for shaping optimal decision boundaries is not new (**Supplementary Table 1**; refs. 24–27). In this work, we provide a general and practical approach, available as an open-source software tool with documentation (http://www.bioconductor.org/packages/IHW and **Supplementary Software**). Most importantly, we show how to establish type I error control and stability, thus overcoming two major limitations of previous approaches.

Building on our preliminary list of suitable covariates for applications (**Table 1**), further work could establish additional domain-specific covariates, formalize and automate the assessment of diagnostic plots and extend IHW to higher-dimensional covariates.

Various approaches for increasing power compared with that of the BH method have focused on estimating the fraction of true nulls among all hypotheses instead of conservatively bounding it by 1 as the BH method does[2]. In practice, this tends to have limited impact, since in the most interesting situations the number of true alternatives is small compared with all tests, and no substantial power increase is gained. On the other hand, such an extension could be beneficial for IHW, since the groups that get assigned a high weight often also have a reduced proportion of true nulls.

The issue of dependence between hypotheses deserves attention. For example, the BH method proof was initially provided under the assumption of independent hypothesis tests and later extended to positive regression dependence[28]. Beyond that, BH has turned out to be remarkably robust to correlations encountered in analyses of real data. In our experience, IHW inherits this property of BH whenever the covariate is not involved in the joint dependence of the null *P*-values.

In our method we have explicitly avoided estimating the densities in equation (1). Nevertheless, the local tdr is an interesting quantity in its own right, since it provides a posterior probability for each individual hypothesis. Our weighted *P*-values do not provide this information. Thus, development of stable estimation procedures for the tdr that incorporate informative covariates is needed and would be complementary to our work[19,22–24].

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank B. Fischer, R. Gentleman, M.I. Love, M. Savitski, O. Stegle, and B. Velten for insightful discussions and comments on the manuscript; the COIN-OR project for the open-source SYMPHONY software and V. Kim for interfacing it to R through the lpsymphony package. We acknowledge support from the European Commission through the Horizon 2020 project SOUND.

AUTHOR CONTRIBUTIONS

N.I. and W.H. developed the method and wrote the manuscript. N.I. implemented the method and performed the analyses. J.Z. analyzed the hQTL data set. B.K. contributed statistical concepts and ideas. All authors read and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Benjamini, Y. & Hochberg, Y. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
2. Benjamini, Y., Krieger, A.M. & Yekutieli, D. *Biometrika* **93**, 491–507 (2006).
3. Storey, J.D., Taylor, J.E. & Siegmund, D. *J. R. Stat. Soc. Series B Stat. Methodol.* **66**, 187–205 (2004).
4. Efron, B. *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Cambridge University Press, 2010).
5. Strimmer, K. *BMC Bioinformatics* **9**, 303 (2008).
6. Genovese, C.R., Roeder, K. & Wasserman, L. *Biometrika* **93**, 509–524 (2006).
7. Roeder, K., Devlin, B. & Wasserman, L. *Genet. Epidemiol.* **31**, 741–747 (2007).
8. Roquain, E. & van de Wiel, M. *Electron. J. Stat.* **3**, 678–711 (2009).
9. Bourgon, R., Gentleman, R. & Huber, W. *Proc. Natl. Acad. Sci. USA* **107**, 9546–9551 (2010).
10. Hu, J.X., Zhao, H. & Zhou, H.H. *J. Am. Stat. Assoc.* **105**, 1215–1227 (2010).
11. Dobriban, E., Fortney, K., Kim, S.K. & Owen, A.B. *Biometrika* **102**, 753–766 (2015).
12. Love, M.I., Huber, W. & Anders, S. *Genome Biol.* **15**, 550 (2014).
13. Bottomly, D. *et al. PLoS One* **6**, e17820 (2011).
14. Frazee, A.C., Langmead, B. & Leek, J.T. *BMC Bioinformatics* **12**, 449 (2011).
15. Dephoure, N. & Gygi, S.P. *Sci. Signal.* **5**, rs2 (2012).
16. Grubert, F. *et al. Cell* **162**, 1051–1065 (2015).
17. Peña, E.A., Habiger, J.D. & Wu, W. *Ann. Stat.* **39**, 556–583 (2011).
18. Sun, W. & Cai, T.T. *J. Am. Stat. Assoc.* **102**, 901–912 (2007).
19. Stephens, M. Preprint at http://biorxiv.org/content/early/2016/01/29/038216.article-info (2016).
20. Cai, T.T. & Sun, W. *J. Am. Stat. Assoc.* **104**, 1467–1481 (2009).
21. Ochoa, A., Storey, J.D., Llinás, M. & Singh, M. *PLoS Comput. Biol.* **11**, e1004509 (2015).
22. Ploner, A., Calza, S., Gusnanto, A. & Pawitan, Y. *Bioinformatics* **22**, 556–565 (2006).
23. Scott, J.G., Kelly, R.C., Smith, M.A., Zhou, P. & Kass, R.E. *J. Am. Stat. Assoc.* **110**, 459–471 (2015).
24. Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G. & Kong, A. *Ann. Appl. Stat.* **2**, 714–735 (2008).
25. Efron, B. & Zhang, N.R. *Biometrika* **98**, 251–271 (2011).
26. Du, L. & Zhang, C. *Ann. Stat.* **42**, 1262–1311 (2014).
27. Yoo, Y.J., Bull, S.B., Paterson, A.D., Waggott, D. & Sun, L. *Genet. Epidemiol.* **34**, 107–118 (2010).
28. Benjamini, Y. & Yekutieli, D. *Ann. Stat.* **29**, 1165–1188 (2001).

ONLINE METHODS

Description of the IHW algorithm. Hypothesis tests in a multiple-testing scenario are divided into G different groups based on the covariate, typically of about equal size. Each group g is associated with weight w_g. The following optimization problem is solved: find the weight vector $w = (w_1, \ldots, w_G)$ that maximizes the number of rejections of the weighted BH method at level α. This method, naive IHW, is modified by the following three extensions.

E1. Instead of the above optimization task, we solve a convex relaxation of it. In statistical terms this corresponds to replacing the empirical cumulative distribution functions (ECDF) of the P-values with the Grenander estimators (least concave majorant of the ECDF). The resulting problem is convex and can be efficiently solved even for large numbers of hypotheses.

E2. We randomly split the hypotheses into k folds. For each fold, we apply convex IHW to the other $k - 1$ folds and assign the learned weights to the remaining fold. Thus the weight assigned to a given hypothesis does not directly depend on its P-value, but only on its covariate.

E3. The performance of the algorithm can be further improved by ensuring that the weights learned with $k - 1$ folds generalize to the held-out fold. Therefore, we introduce a regularization parameter $\lambda \geq 0$, and the optimization is done over a constrained subset of the weights. For an ordered covariate, we require that

$$\sum_{g=2}^{G} |w_g - w_{g-1}| \leq \lambda$$

i.e., weights of successive groups should not be too different. For an unordered covariate, we use instead the constraint

$$\sum_{g=1}^{G} |w_g - 1| \leq \lambda$$

i.e., deviations from 1 are penalized. In the limit case $\lambda = 0$, all weights are the same, so IHW with $\lambda = 0$ is just the BH method. IHW with $\lambda \to \infty$ is the unconstrained version. Choice of λ is a model selection problem, so within each split in E2 we apply a second nested layer of cross-validation. E3 is optional; whether or not to apply it will depend on the data. It will be most beneficial if the number of hypotheses per group is relatively small.

A complete description of the algorithm, including an efficient computational implementation of the optimization task, is provided in **Supplementary Note 2**. **Supplementary Note 7** describes its theoretical justification.

Code availability. The IHW package is available from Bioconductor at http://www.bioconductor.org/packages/IHW. It comes with detailed documentation and a vignette that showcases the application of IHW to a real data set. The vignette also provides guidance on the choice of informative covariates and suggests diagnostic plots so that users can determine whether their covariate satisfies the required conditions.

Executable documents (Rmarkdown) reproducing all analyses shown here can be downloaded at http://bioconductor.org/packages/IHWpaper.