Published in final edited form as:
Nat Methods. 2016 July ; 13(7): 577–580. doi:10.1038/nmeth.3885.

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

Nikolaos Ignatiadis, **Bernd Klaus**, **Judith Zaugg**, and **Wolfgang Huber**[1]
European Molecular Biology Laboratory, Heidelberg, Germany

Abstract

Hypothesis weighting improves the power of large-scale multiple testing. We describe a method that uses covariates independent of the p-values under the null hypothesis, but informative of each test's power or prior probability of the null hypothesis. Independent hypothesis weighting (IHW) increases power while controlling the false discovery rate (FDR). IHW is a practical approach to discover associations in large datasets as encountered in genomics and high-throughput biology. Availability: www.bioconductor.org/packages/IHW

Multiple 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 FDR, i. e., the expected fraction of false discoveries. Commonly used procedures, such as that of Benjamini and Hochberg, achieve this objective by working solely off the list of p-values [1–5]. However, such an approach has suboptimal power when the individual tests differ in their statistical properties, such as sample size, true effect size, signal-to-noise ratio, or prior probability of being false.

For example, in RNA-seq differential expression analysis, each test is associated with a different gene, and because of differences in the number of reads mapped the genes greatly differ in their signal-to-noise ratio. In genome-wise association studies (GWAS), associations are sought between genetic polymorphisms and phenotypic traits; however, the power to detect an association is lower for rarer polymorphisms (all else being equal). In GWAS of gene expression phenotypes (eQTL), 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, a non-negative number as a measure of its priority (Supplementary Note 1). The weights fulfill a budget criterion, commonly that they average to one. Hypotheses with higher weights get prioritized [6]. The procedure of Benjamini and

[1] Corresponding Author: whuber@embl.de.

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 dataset. 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.

Hochberg (BH) [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 controls the FDR if the weights are pre-specified and thus independent of the data. However, the optimal choice of weights is rarely known in practice, and a generally applicable data-driven method would be desirable [7–11].

Independent hypothesis weighting (IHW) is a multiple testing procedure that applies the weighted BH method [6] using weights derived from the data. The input to IHW is a two-column table of p-values and covariates. The covariate can be any continuous-valued or categorical variable that is thought to be informative on the statistical properties of the hypothesis tests, while it is 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 (Fig. 1).

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 step-wise 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.

An initial implementation ("naive IHW") is easy to explain. The algorithm divides the tests into groups based on the covariate. Then, we associate each group 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.

In many applications, this approach is already satisfactory, but it has two shortcomings: First, the underlying optimization problem is difficult and does not easily scale to problems with millions of tests. Second, in certain situations, described below, this algorithm leads to loss of type I error control. The reason for the latter is analogous to overfitting in statistical learning, and we use methods from this field to overcome the shortcomings: convex relaxation, data splitting and regularization (Online Methods and Supplementary Note 2). The full IHW algorithm employs these three extensions.

IHW increases empirical detection power compared to the BH procedure. We illustrate this claim on three exemplary applications (Supplementary Note 3). The first, by Bottomly *et al.* [13, 14], is an RNA-seq dataset used to detect differential gene expression between mouse strains. p-values were calculated using DESeq2 [12]. Here we used the mean of normalized counts for each gene, across samples, as the informative covariate. We saw an increased number of discoveries compared to BH (Fig. 2a). In addition, we observed that the learned weight function prioritized genes with higher mean normalized counts (Supplementary Fig. 1a).

Second, we analyzed a quantitative mass-spectrometry (SILAC) experiment in which yeast cells treated with rapamycin were compared to yeast treated with DMSO (2 × 6 biological replicates) [15]. Differential protein abundance of 2,666 proteins was evaluated using Welch's *t*-test [15]. As a covariate we used the total number of peptides that were quantified across all samples for each protein. IHW again showed increased power compared to 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 modification marks (H3K27ac) [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 to 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, it had the advantage that no arbitrary choice of distance threshold was needed, and that the weights were more nuanced than a hard distance threshold. IHW does not exclude SNP-phenotype pairs far away, and these can still be detected as long as they have a sufficiently small p-value.

The extensions to naive IHW are needed to ensure type I error control. Naive IHW, as well as previous approaches to data-driven hypothesis weighting or filtering, 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, the local fdr methods (Clfdr and FDRreg) often show strong deviations from the target FDR in a direction (conservative or anti-conservative) 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. The results of our method comparisons are summarized in Table 2 (Fig. 2 and Supplementary Fig. 2), and the simulations are described in Supplementary Note 4.

IHW can apply a size investing strategy. IHW assigns low weight to covariate-groups with low signal (such as Fig. 1d). While this may be expected, a less intuitive effect can pertain to groups with very small p-values. IHW can move away weight from these towards groups with more intermediate p-values, since the former will be rejected even with a lower weight. This is called *size investing* [17]. Several other methods (Table 2), including Independent Filtering, stratified BH, LSL-GBH and FDRreg, cannot apply size investing and might even lose power compared to the BH method (Supplementary Fig. 2d,f and Supplementary Note 5).

It is instructive to consider the relation between IHW and the concept of local true discovery rates. p-values are a reduction of data into one number, which typically does not contain all the important information (Table 1; [18, 19]). One might wonder whether there are other quantities that are better suited for selecting discoveries. The theoretically optimal candidate is the *local true discovery rate* (tdr) [4]. The tdr of the i^{th} hypothesis is [4]

$$\mathrm{tdr}_{\mathrm{i}}(p)=\pi_{1,\mathrm{i}}\frac{\mathrm{f}_{1,i}(p)}{\mathrm{f}_{\mathrm{i}}(p)} \quad (1)$$

A schematic explanation is given in Fig. 3a (see also Supplementary Figs. 3 and 4). f_i is the density of the distribution of the p-value *p*. It is a mixture of two distributions, $f_i = \pi_{0,i} f_0 + \pi_{1,i} f_{1,i}$, where the densities f_0 and $f_{1,i}$ are conditional on the null or the alternative 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*. In Fig. 3b-d three hypotheses are shown with different tdr curves corresponding to different power profiles.

It can now be shown that to maximize power at a given FDR, one should reject the hypotheses with the highest tdr [20, 21]. In other words, if we knew the functions in Equation (1) and could use 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 can now assume that the many different, unknown univariate functions tdr_i (p), one for each hypothesis *i*, can be approximated by a single bivariate function tdr(*p, x*), where *x* is the covariate. The joint density of *p* and *x* (Fig. 3e) gives rise to the joint density of tdr and *x* (Fig. 3f). We can see how in such a scenario the decision boundary of the BH method tends to be suboptimal. As it is defined solely in terms of p-values (Fig. 3e), it differs from the optimal region, whose boundary is a vertical line of constant tdr (Fig. 3f).

However, in practice, 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.

With IHW we circumvent explicit estimation of the bivariate tdr function and instead derive a powerful testing procedure by assigning data-driven hypothesis weights. In addition, the IHW method readily extends to other weighted multiple testing procedures [6]. In Supplementary Note 6 (and Supplementary Fig. 5) we describe IHW-Bonferroni, a new powerful method for control of 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. It 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 (Table 2; [24–26]). In this work, we provide a general and practical approach. Most importantly, we show how to overcome two major limitations of previous approaches: type I error control and stability.

We gave examples of suitable covariates for a variety of applications in Table 1. Further work could establish additional domain-specific choices of covariates, formalize and automate the assessment of diagnostic plots such as Fig. 1 and extend IHW to higher dimensional covariates.

Various approaches to increasing power compared to the BH method have focused on estimating the fraction of true nulls among all hypotheses instead of conservatively approximating 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 discoveries is small compared to all tests and no substantial power increase is gained. On the other hand, such an extension could be beneficial for IHW, since often the groups that get assigned a high weight also have a reduced proportion of true nulls.

The issue of dependence between hypotheses deserves attention. For example, the proof of the BH method was initially provided under the assumption of independent hypothesis tests and later extended to positive regression dependence [27]. 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 true discovery rate 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 local local true discovery rate that incorporate informative covariates is needed and would be complementary to our work [19, 22–24].

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 dataset. The vignette also provides guidance on the choice of informative covariates and suggests diagnostic plots, so that users can determine if their covariate satisfies the required conditions.

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

Both packages are also available as Supplementary Software to this manuscript.

Online Methods

Description of the IHW algorithm

The hypothesis tests 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 $\boldsymbol{w} = (w_1, \ldots, w_G)$ 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank B. Fischer, M. I. Love, M. Savitski and O. Stegle 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.

References

[1]. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Statistical Methodology). 1995:289–300.

[2]. Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. Biometrika. 2006; 93:491–507.

[3]. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2004; 66:187–205.

[4]. Efron, B. Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction. Cambridge University Press; 2010.

[5]. Strimmer K. A unified approach to false discovery rate estimation. BMC Bioinformatics. 2008; 9:303. [PubMed: 18613966]

[6]. Genovese CR, Roeder K, Wasserman L. False discovery control with p-value weighting. Biometrika. 2006; 93:509–524.

[7]. Roeder K, Devlin B, Wasserman L. Improving power in genome-wide association studies: weights tip the scale. Genetic Epidemiology. 2007; 31:741–747. [PubMed: 17549760]

[8]. Roquain E, Van De Wiel M. Optimal weighting for false discovery rate control. Electronic Journal of Statistics. 2009; 3:678–711.

[9]. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. Proceedings of the National Academy of Sciences. 2010; 107:9546–9551.

[10]. Hu JX, Zhao H, Zhou HH. False discovery rate control with groups. Journal of the American Statistical Association. 2010; 105

[11]. Dobriban E, Fortney K, Kim SK, Owen AB. Optimal multiple testing under a Gaussian prior on the effect sizes. Biometrika. 2015; 102:753–766. [PubMed: 27046938]

[12]. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biology. 2014; 15:550. [PubMed: 25516281]

[13]. Bottomly D, et al. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. PLoS ONE. 2011; 6:e17820. [PubMed: 21455293]

[14]. Frazee AC, Langmead B, Leek JT. Recount: a multi-experiment resource of analysis-ready rna-seq gene count datasets. BMC Bioinformatics. 2011; 12:449. [PubMed: 22087737]

[15]. Dephoure N, Gygi SP. Hyperplexing: a method for higher-order multiplexed quantitative proteomics provides a map of the dynamic response to rapamycin in yeast. Science Signaling. 2012; 5:rs2–rs2. [PubMed: 22457332]

[16]. Grubert F, et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. Cell. 2015; 162:1051–1065. [PubMed: 26300125]

[17]. Peña EA, Habiger JD, Wu W. Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. The Annals of Statistics. 2011; 39:556–583. [PubMed: 25018568]

[18]. Sun W, Cai TT. Oracle and adaptive compound decision rules for false discovery rate control. Journal of the American Statistical Association. 2007; 102:901–912.

[19]. Stephens M. False discovery rates: A new deal. bioRxiv. 2016:038216.

[20]. Cai TT, Sun W. Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. Journal of the American Statistical Association. 2009; 104

[21]. Ochoa A, Storey JD, Llins M, Singh M. Beyond the E-value: Stratified statistics for protein domain prediction. PLoS Computational Biology. 2015; 11:e1004509. [PubMed: 26575353]

[22]. Ploner A, Calza S, Gusnanto A, Pawitan Y. Multidimensional local false discovery rate for microarray studies. Bioinformatics. 2006; 22:556–565. [PubMed: 16368770]

[23]. Scott JG, Kelly RC, Smith MA, Zhou P, Kass RE. False discovery rate regression: an application to neural synchrony detection in primary visual cortex. Journal of the American Statistical Association. 2015; 110:459–471. [PubMed: 26855459]

[24]. Ferkingstad E, Frigessi A, Rue H, Thorleifsson G, Kong A. Unsupervised empirical Bayesian multiple testing with external covariates. The Annals of Applied Statistics. 2008:714–735.

[25]. Efron B, Zhang NR. False discovery rates and copy number variation. Biometrika. 2011; 98:251–271.

[26]. Du L, Zhang C. Single-index modulated multiple testing. The Annals of Statistics. 2014; 42:30–79.

[27]. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics. 2001:1165–1188.

[28]. Yoo YJ, Bull SB, Paterson AD, Waggott D, Sun L. Were genome-wide linkage studies a waste of time? Exploiting candidate regions within genome-wide association studies. Genetic Epidemiology. 2010; 34:107–118. [PubMed: 19626703]

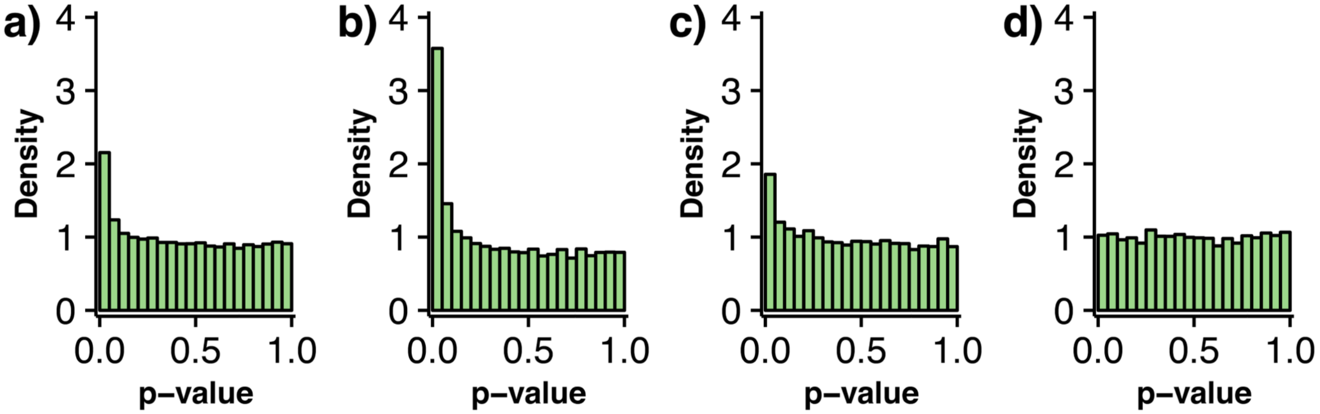

Figure 1. Histograms stratified by the covariate as a diagnostic plot.

a) The histogram of all p-values shows a mixture of a uniform distribution (corresponding to the true null hypotheses) and an enrichment of small p-values to the left (corresponding to the alternatives). 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. Shown is an example of a good covariate: each histogram still shows a uniform component, but the mixture proportion and/or the shape of the alternative distribution differ between the groups. If all histograms look the same, the covariate is uninformative, and its use would 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.

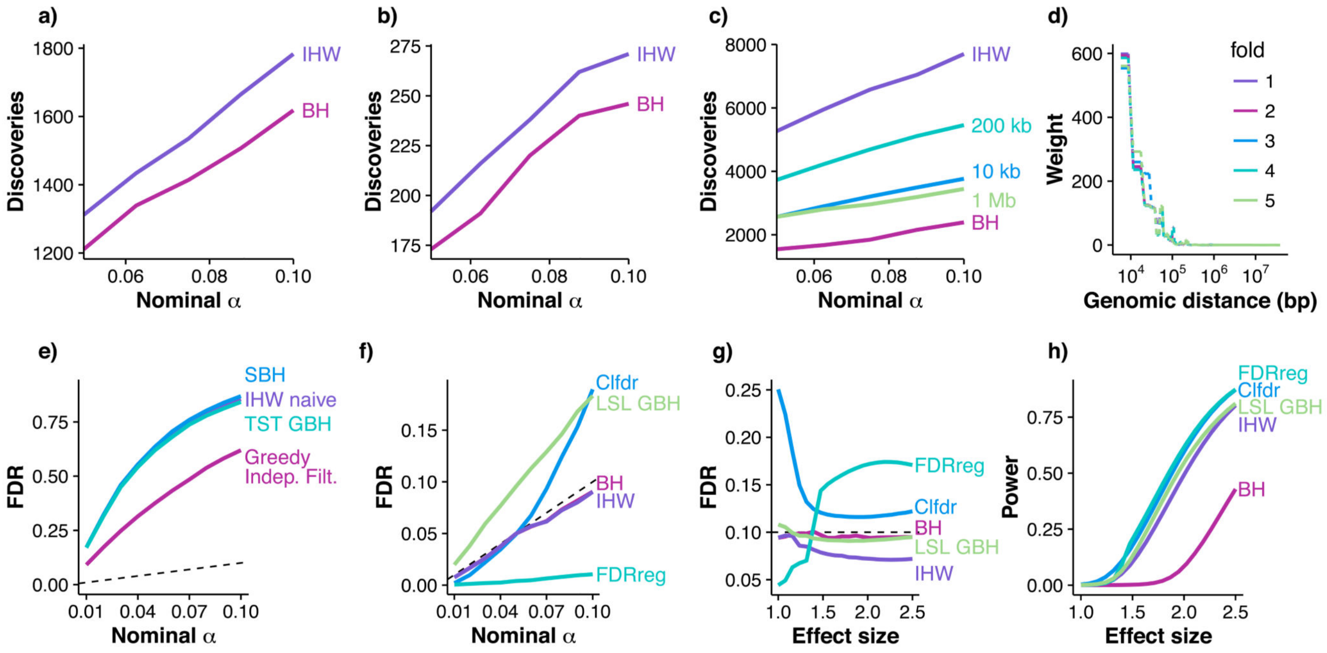

Figure 2. Performance evaluation.

Panels **a-c** show the number of discoveries with IHW and BH on real data as a function of the target FDR. **a)** RNA-Seq dataset [13] with mean of normalized counts for each gene as the covariate. **b)** SILAC dataset [15], with number of peptides quantified per protein as the covariate. **c)** hQTL dataset [16] for Chromosome 21, with genomic distance between SNPs and ChIP-seq signals as the covariate. Independent Filtering with different distance cutoffs was also applied. **d)** Weight function learned by IHW at $\alpha = 0.1$ for the hQTL dataset. Shown are the curves for the five folds in the data splitting scheme. Panels **e-h** benchmark different methods based on simulations. Brief descriptions of each method are in Table 2. **e–f)** Type I error control if all null hypotheses are true. Shown is the true FDR against the nominal significance level α. **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. **h)** Power analysis. All methods show improvement over BH.

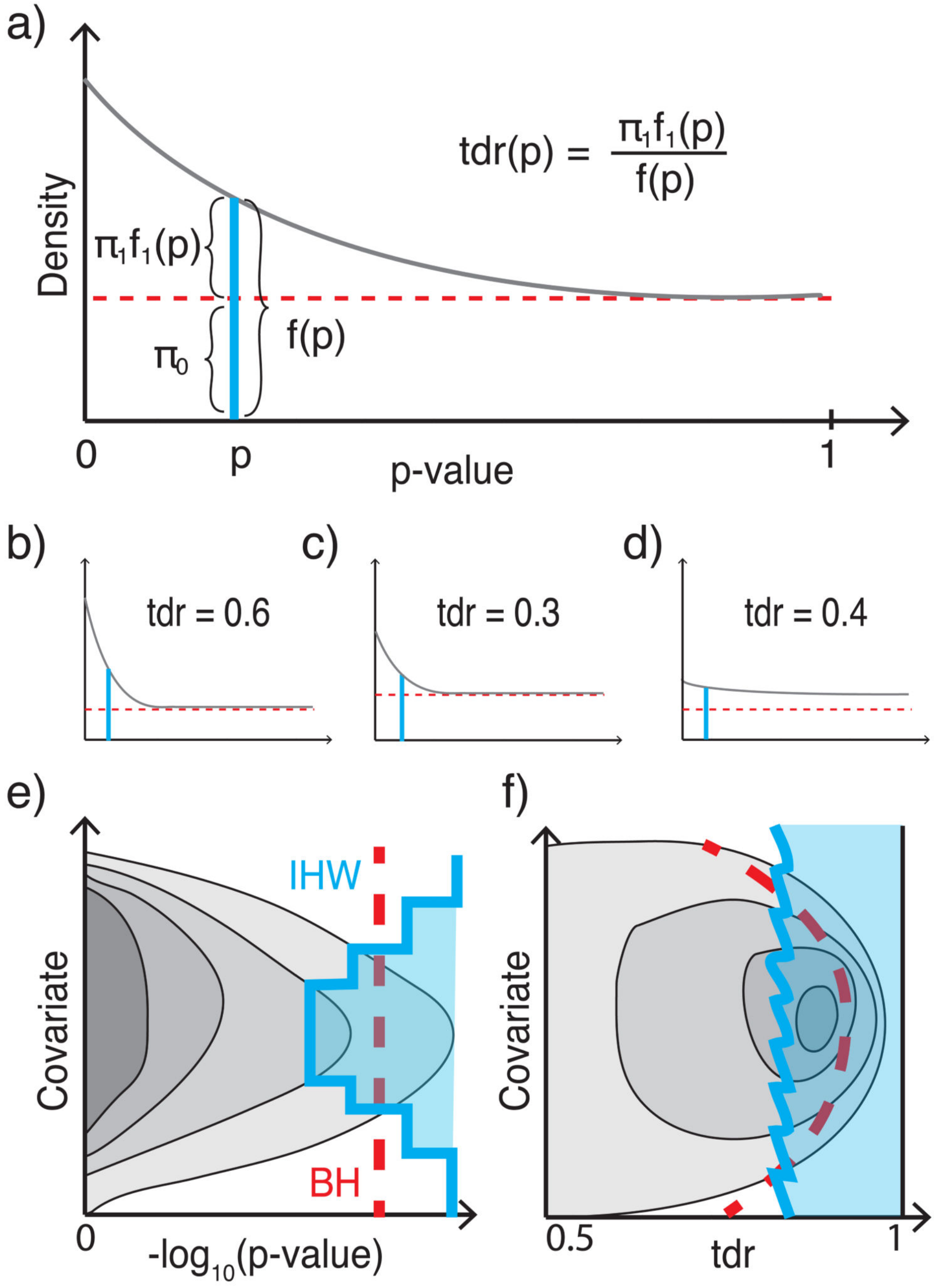

Figure 3. True discovery rate and informative covariates.

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 true discovery rate (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 like 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**) By 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 too low otherwise. In contrast, IHW approximates a line of constant tdr, implying efficient use of the FDR budget. An important feature of IHW is that it works directly on p-values and covariates rather than explicitly estimating the tdr.

Table 1

Examples of covariates.

Application	**Covariate**
Differential expression analysis	Sum of read counts per gene across all samples [12]
Genome-wide association study (GWAS)	Minor allele frequency
Expression-QTL analysis	Distance between the genetic variant and genomic location of the phenotype
ChIP-QTL analysis	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

Table 2

Short description of the different methods benchmarked and summary of the results of Fig. 2e–h and Supplementary Fig. 2.

Method	Short description	Type I error control		Gain in power		Comment
		π_0=1	t-test	t-test (vs BH)	size investing	
BH	Method of Benjamini and Hochberg [1] to control false-discovery rate (FDR) for multiple exchangeable hypotheses.	Yes	Yes	–	–	
IHW	Independent hypothesis weighting, as proposed here.	Yes	Yes	Yes	Yes	
Naive IHW	Naive independent hypothesis weighting, as proposed here.	No	No	Yes	Yes	
Greedy Independent Filtering	The Independent Filtering procedure [9] modified to use a data-driven filter threshold which maximizes the number of discoveries.	No	No	Yes	No	The covariate-weights function is a binary step, monotonic.
SBH	Stratified Benjamini-Hochberg [28]: Apply the BH procedure at level α within each stratum, then combine the discoveries across the strata.	No	No	Yes	No	
TST-GBH	The Group BH procedure [10]: An adaptive weighted BH procedure applied with weights proportional to π_1/π_0 within each group. π_0 is estimated using the TST estimator [2].	No	No	Yes	No	
LSL-GBH	The Group BH procedure [10], where π_0 is estimated using the LSL estimator	Yes	Yes	Yes	No	
Clfdr	In the Clfdr procedure [20], the local fdr is estimated separately within each group and the estimates are pooled together. For the fdr estimation here we use the modified Grenander estimator [5].	Yes	No	Yes	Yes	
FDRreg	The FDR regression method [23] estimates the local fdr by assuming all hypotheses have the same alternative density and π_0 varies smoothly as a function of the covariate.	Yes	No	Yes	No	Requires *z*-scores.