PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Extracting replicable associations across multiple studies: Empirical Bayes algorithms for controlling the false discovery rate

David Amar[1], **Ron Shamir**[1]*, **Daniel Yekutieli**[2]

1 The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel, **2** Department of Statistics and OR, Tel Aviv University, Tel Aviv, Israel

* rshamir@tau.ac.il

Citation: Amar D, Shamir R, Yekutieli D (2017) Extracting replicable associations across multiple studies: Empirical Bayes algorithms for controlling the false discovery rate. PLoS Comput Biol 13(8): e1005700. https://doi.org/10.1371/journal.pcbi.1005700

Editor: Alon Keinan, Cornell University, UNITED STATES

Received: February 9, 2017

Accepted: July 24, 2017

Published: August 18, 2017

Data Availability Statement: The datasets and an implementation of SCREEN are available in supporting information file S1 Data. In addition, the sources for the data are referenced in the manuscript.

Funding: This research was supported in part by the Israel Science Foundation as part of the ISF-NSFC joint program, by the Raymond and Beverly Sackler Chair in Bioinformatics, and by the Israeli Center of Research Excellence (I-CORE), Gene Regulation in Complex Human Disease, Center No

Abstract

In almost every field in genomics, large-scale biomedical datasets are used to report associations. Extracting associations that recur across multiple studies while controlling the false discovery rate is a fundamental challenge. Here, we propose a new method to allow joint analysis of multiple studies. Given a set of p-values obtained from each study, the goal is to identify associations that recur in at least $k > 1$ studies while controlling the false discovery rate. We propose several new algorithms that differ in how the study dependencies are modeled, and compare them and extant methods under various simulated scenarios. The top algorithm, SCREEN (Scalable Cluster-based REplicability ENhancement), is our new algorithm that works in three stages: (1) clustering an estimated correlation network of the studies, (2) learning replicability (e.g., of genes) within clusters, and (3) merging the results across the clusters. When we applied SCREEN to two real datasets it greatly outperformed the results obtained via standard meta-analysis. First, on a collection of 29 case-control gene expression cancer studies, we detected a large set of consistently up-regulated genes related to proliferation and cell cycle regulation. These genes are both consistently up-regulated across many cancer studies, and are well connected in known gene networks. Second, on a recent pan-cancer study that examined the expression profiles of patients with and without mutations in the HLA complex, we detected a large active module of up-regulated genes that are both related to immune responses and are well connected in known gene networks. This module covers thrice more genes as compared to the original study at a similar false discovery rate, demonstrating the high power of SCREEN. An implementation of SCREEN is available in the supplement.

Author summary

When analyzing results from multiple studies, extracting replicated associations is the first step towards making new discoveries. The standard approach for this task is to use meta-analysis methods, which usually make an underlying null hypothesis that a gene has no effect in all studies. On the other hand, in replicability analysis we explicitly require that

41/11. Part of the work was done while DA and RS were visiting the Simons Institute for the Theory of Computing. The work of DY was supported by ISF grant 1049/16. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

the gene will manifest a recurring pattern of effects. In this study we develop new algorithms for replicability analysis that are both scalable (i.e., can handle many studies) and allow controlling the false discovery rate. We show that our main algorithm called SCREEN (Scalable Cluster-based REplicability ENhancement) outperforms the other methods in simulated scenarios. Moreover, when applied to real datasets, SCREEN greatly extended the results of the meta-analysis, and can even facilitate detection of new biological results.

Introduction

Confidence in reported findings is a prerequisite for advancing any scientific field. Such confidence is achieved by showing replication of discoveries in new studies [1]. In recent years studies have shown low reproducibility of results in several domains, including economics [2], psychology [3], medicine [4], and biology [5–7]. A new methodology called replicability analysis was recently suggested as a way to statistically pinpoint replicated discoveries across studies while controlling for the false discovery rate (FDR) [8]. This type of analysis is essential when trying to detect new hypotheses by integration of existing data from multiple high-throughput experiments.

The practical importance of replicability analysis is twofold. First, it quantifies the reliability of reported results. Second, collated information from multiple studies can identify scientific results that are beyond the reach of each single study. Indeed, in Genome Wide Association Studies (GWAS) replicability analysis allowed detection of new results that were not identified in meta-analysis, demonstrating that the two approaches are complementary [9].

Meta-analyses are widely applied and have been extensively studied in statistics [10] and in computational biology [11, 12]. However, in recent years the changes in the scale and scope of public high-throughput biomedical data have posed new methodological challenges. The first, and more obvious, is accounting for inflation in the number of false discoveries due to the multiplicity of outcomes, as hundreds of thousands and even millions of hypotheses are tested (see Zeggini et al. [13] for example). The second challenge is directly assessing consistency of results, which is not addressed by the classic null hypothesis of meta-analysis that the effect size is 0 in all the studies. Third, there is a need to distinguish between true effects that are specific to a single study and true effects that represent general discoveries and thus are replicable. For example, Kraft et al. [14] suggested that the effect of common genetic variants on the phenotype may correlate with population biases in a specific GWAS. While these are real discoveries in the sense that similar estimated effects are expected to be observed if the experiment could be replicated on the same cohort, their scientific importance is limited because they are specific to that cohort. For this reason, the authors argue that it is important to identify the association in additional studies conducted using a similar, but not identical, study base.

In recent years several frequentist approaches were suggested for the problem. Benjamini and Heller [15] introduced an inferential framework for replicability that is based on tests of partial conjunction null hypotheses. For meta-analysis of n studies of the same m outcomes and $u = 1...n$, the partial conjunction $H^{u/n}(g)$ is that outcome g has a non-null effect in less than u studies. Thus $H^{1/n}(g)$ is the standard meta-analysis null hypothesis that outcome g has a null effect in all n studies. The authors introduced p-values for testing $H^{u/n}(g)$ for each outcome. Benjamini, Heller and Yekutieli [8] applied the Benjamini-Hochberg FDR procedure [16] (BH) to the partial conjunction hypotheses p-values, and suggested setting $u = 2$ in order to assess replicability. Heller et al. [17] developed an approach for checking if a follow-up study corroborates the results reported in the original study. Song and Tseng [18] proposed a method to evaluate the proportion of non-null effects of a gene. However, they used the standard meta-analysis null hypothesis and their method cannot handle composite hypotheses, which are partial conjunctions $H^{u/n}$ with $u > 1$.

Bayesian methods handle these shortcomings and offer a powerful framework for replicability analysis. For analyzing results from a single study, Efron introduced an empirical Bayes framework called the two-groups model [19]. It allows explicit analysis of the distribution of the statistic (e.g., p-values) of the underlying null and non-null groups. This clustering-based structure is then used to quantify the FDR of a rejection rule, and to compute a single point statistic, which is referred to as local Bayes FDR, or simply *fdr*. Heller and Yekutieli [9] introduced a method called repfdr, which extends the two-groups model for testing the partial conjunction hypotheses to the multi-study case. Formally, the problem is as follows: given an $n \times m$ matrix Z, where $Z_{i,j}$ is the p-value (or z-score) of object (e.g., gene) i in study j, our goal is to identify the objects that are k-replicable (i.e., significant in k or more studies) while controlling the fdr.

Repfdr estimates the posterior probabilities of the various configurations of outcome effect status (null or non-null) across studies, and computes the fdr for each partial conjunction null by summing the posterior probabilities for the relevant configurations. The authors showed that their approach controls the FDR and offers more power than the frequentist methods. However, repfdr is not scalable and can only handle a few datasets. In addition, it was particularly designed for GWAS datasets in which the number of tested objects (e.g., SNPs) is very large (i.e., > 100k).

In this study, we propose ways to overcome the limitations of repfdr. Our focus is on allowing efficient computation when m is large and n is limited, such as in gene expression datasets (i.e., $n \sim 20$k). To reach this goal we make three simplifying assumptions: (1) we ignore the effect size, (2) we ignore the direction of the statistic, and (3) we assume that the studies originate from independent clusters. In addition, to handle larger values of m we compute an upper bound for the *fdr*, cutting the running time substantially. Our main algorithm is called SCREEN (Scalable Cluster-based REplicability ENhancement). It first detects the study clusters, then uses Expectation-Maximization (EM) to model each cluster, and finally merges the clusters using dynamic programming. Other algorithms that we propose here include two variants of SCREEN that differ in the way the studies are clustered: *SCREEN–ind* assumes independence and treats each study as a single cluster, and repfdr-UB puts all studies in one cluster. We compared SCREEN to other algorithms using various simulated scenarios and showed that only SCREEN had consistently low empirical false discovery proportions, and very high detection power.

We applied SCREEN to two cancer datasets, where each is a collection of case-control gene expression experiments. In both cases SCREEN greatly improved the results obtained by standard meta-analysis, and provided new biological insights. The first dataset is a collection of 29 case-control gene expression cancer studies from different tissues. Here, SCREEN detected a large set of genes that are consistently up-regulated, highly enriched for cell proliferation and cell cycle regulation functions, and are well connected in known gene networks, indicating their functional coherence. The second dataset is a recent pan-cancer study that examined the expression profiles of patients with and without mutations in the HLA complex across 11 cancer types [20]. SCREEN detected a large set of up-regulated genes that are related to immune responses. Importantly, SCREEN reported many more immune response genes than the original study thanks to our ability to quantify the fdr, and allowed detection of prominent genes and pathways that were not reported previously.

Results

Outline: After introducing some background and notation, we present a dynamic programming algorithm for calculating the fdr under the assumption that the studies are independent. Second, we discuss EM-based algorithms for dealing with dependence. We then show that given the prior probability of only a subset of all configurations an upper-bound for the *fdr* can be computed. This leads us to a much faster algorithm that can handle many studies. Third, we extend the dynamic programming algorithm to handle independent clusters of studies. This will lead us to the complete SCREEN algorithm, which is based on EM-based dependency modeling within study clusters and merging the results using dynamic programming. Finally, we discuss experimental results on simulated and real datasets.

Preliminaries and notations

We start with a brief introduction to the single-study model. For a full description and background see [19]. Given a large set of N hypotheses tested in a large-scale study, the two-groups model provides a simple Bayesian framework for multiple testing: each of the N cases (e.g., genes in a gene expression study) are either null or non-null with prior probability π_0 and $\pi_1 = 1 - \pi_0$, and with z-scores (or p-values) having density either $f_0(z)$ or $f_1(z)$. When the assumptions of the statistical test are valid, we know that the f_0 distribution is a standard normal (or a uniform distribution for p-values), and we call it the theoretical null. The mixture density and probability distributions are:

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$$
$$F(z) = \pi_0 F_0(z) + \pi_1 F_1(z)$$

For a rejection area $\mathcal{Z}_y = (-\infty, y)$, using Bayes rule we get:

$$Fdr(\mathcal{Z}_y) \equiv Pr\{null|z \in \mathcal{Z}_y\} = \pi_0 F_0(y)/F(y)$$

We call *Fdr* the (Bayes) false discovery rate for $\mathcal{Z}$: this is the probability we would make a false discovery if we report $\mathcal{Z}$ as non-null. If $\mathcal{Z}$ is a single point z_0 we define the local (Bayes) false discovery rate as:

$$fdr(z_0) \equiv Pr\{null|z = z_0\} = \pi_0 f_0(z_0)/f(z_0)$$

Previous work have shown that: (1) the Bayes Fdr is tightly related to false discovery control in the frequentist sense, and (2) using a threshold on the local fdr for defining discoveries is equivalent to the optimal Bayes rule in terms of classification between nulls and non-nulls. Moreover, a threshold of 0.2 on the local fdr was suggested [19, 21]. Note that in our Bayesian setting computing the local fdr of a gene is an estimation problem. Within the context of our study we incorporated and tested two established methods for two-groups estimation studies: locfdr [22] and estimation based on mixture of Gaussians, which we call normix, see Materials and methods for details.

Consider an extension of the two group model to analysis of n genes over $m > 1$ studies. The data for gene i are a vector of m statistics $Z_{i,\cdot} = (Z_{i,1}, \cdots, Z_{i,m})$ that are all either z-scores or p-values. For simplicity, from now on we assume that these data are z-scores. The unknown parameter for gene $i = 1, \cdots, n$ is a binary configuration vector $H_{i,\cdot} = (H_{i,1}, \cdots, H_{i,m})$, with $H_{i,j} \in \{0, 1\}$. If $H_{i,j} = 0$ then gene i is a *null* realization in study j, and it is a *non–null* realization otherwise.

We assume that in each study j the parameters of the two-groups model $\theta_j : (\pi_0^j, f_0^j, f_1^j, f^j)$ are fixed and focus on replicability analysis. Generally, unless mentioned otherwise, we assume that the genes are independent. However, note that estimation of θ_j can account for gene dependence within study j [23, 24]. Finally, we also assume that the z-scores of a gene are independent given its configuration. That is,

$$P(Z_{i,\cdot}|H_{i,\cdot}) = \prod_{j=1}^{m} P(Z_{i,j}|H_{i,j}) = \prod_{j=1}^{m} (f_0^j(Z_{i,j}))^{(1-H_{i,j})} (f_1^j(Z_{i,j}))^{H_{i,j}}$$

Next, we use $h \in \{0, 1\}^m$ to denote an arbitrary configuration vector, and $\pi(h)$ to denote a probability assigned to the parameter space. We assume that the researcher has a set of configurations $\mathcal{H}_1 \subseteq \{0,1\}^m$ that represents the desired rejected genes. Here we will assume that $\mathcal{H}_1$ corresponds to genes that are non-null in at least k studies: $\mathcal{H}_1 = \{h : |h| \geq k\}$, where $|h| = \sum_{j=1}^{m} h_j$.

As a note, selection of k depends on the research question at hand. For example, Heller and Yekutieli used $k = 2$ to detect minimal replicability of SNPs in a GWAS [9]. Low k values can also be reasonable if the m studies represent different biological questions that are related, such as differential expression experiments from different cancer subtypes. On the other hand, if the m studies represent tightly related experiments such as biological replicates then larger k (e.g., $m/2$) seems more reasonable.

The local false discovery rate (fdr) of a gene i can be formulated as:

$$fdr(Z_{i,\cdot}) = Pr(\bar{\mathcal{H}}_1|Z_{i,\cdot}) = \sum_{h:h\notin\mathcal{H}_1} P(h|Z_{i,\cdot}) = \sum_{h:h\notin\mathcal{H}_1} \frac{P(Z_{i,\cdot}|h)P(h)}{P(Z_{i,\cdot})}$$

For a given k and $\mathcal{H}_1 = \{h : |h| \geq k\}$ we get:

$$fdr_k(Z_{i,\cdot}) = \sum_{h:|h|<k} \frac{P(Z_{i,\cdot}|h)P(h)}{P(Z_{i,\cdot})}$$

An $O(mnk)$ algorithm when studies are independent

We first address the case where studies are independent.

Lemma. If the studies are independent (in the parameter space) then:

$$fdr_k(Z_{i,\cdot}) = \sum_{h:|h|<k} \prod_{j=1}^{m} \frac{P(Z_{i,j}|h_j)P(h_j)}{f^j(Z_{i,j})}$$

Proof. First, note that under the independence assumption $P(h) = \prod_{j=1}^{m} P(h_j) = \prod_{j=1}^{m} \pi_0^{j\,(1-h_j)}(1-\pi_0^j)^{h_j}$. Second, as the z-scores are independent given the configuration vector h we get that:

$$\begin{aligned} P(Z_{i,\cdot}) &= \sum_h P(Z_{i,\cdot}|h)P(h) = \sum_h \prod_{j=1}^{m} P(Z_{i,j}|h_j)\pi_0^{j\,(1-h_j)}(1-\pi_0^j)^{h_j} \\ &= \prod_{j=1}^{m} \left(P(Z_{i,j}|h_j = 0)\pi_0^j + P(Z_{i,j}|h_j = 1)(1-\pi_0^j) \right) \end{aligned}$$

Proposition: If the studies are independent then fdr_k can be computed in $O(mnk)$.

Proof. By the lemma, the *fdr* of a gene is based on the product of the two-group model densities in each study. Therefore:

$$fdr_k^{indep}(Z_{i,\cdot}) = \sum_{h:|h|<k} \prod_{j=1}^{m} \frac{(\pi_0^j f_0^j(Z_{i,j}))^{1-h_j}((1-\pi_0^j)f_1^j(Z_{i,j}))^{h_j}}{f^j(Z_{i,j})}$$

We use dynamic programming to calculate $fdr_k(z_i)$ as follows. Define:

$$U[i,j,k^*] = \sum_{h:|h|=(k^*-1)} \prod_{j=1}^{m} \frac{(\pi_0^j f_0^j(Z_{i,j}))^{1-h_j}((1-\pi_0^j)f_1^j(Z_{i,j}))^{h_j}}{f^j(Z_{i,j})}$$

These values can be calculated (for each gene *i*) by updating a table of $m \times (k + 1)$ values. The base cases are:

$$U[i,j,1] = \prod_{j=1}^{m} \frac{\pi_0^j f_0^j(Z_{i,j})}{f^j(Z_{i,j})}$$

The recursive formulas are:

$$U[i,j,k^*] \quad = \quad \frac{\pi_0^j f_0^j(Z_{i,j})}{f^j(Z_{i,j})} U[i,j-1,k^*] + \frac{(1-\pi_0)^j f_1^j(Z_{i,j})}{f^j(Z_{i,j})} U[i,j-1,k^*-1]$$

Finally, to obtain the *fdr* of a gene we sum over the values in the last column:

$$fdr_k^{indep}(Z_{i,\cdot}) = \sum_{k^*=1}^{k-1} U[i,m,k^*]$$

The running time for analyzing each gene is $O(mk)$ and the total running time is $O(nmk)$.

Schemes for handling dependencies

The empirical Bayes method of [9] estimates the prior distribution $\pi(h)$ directly from the data. This approach has two drawbacks. First, the EM algorithm explicitly keeps a value for each possible configuration, which makes the algorithm intractable when m is large. Second, the estimation for rare configurations might be inaccurate, unless $n >> 2^m$. As an alternative, we develop an algorithm that keeps in memory only a small set of high probability configurations. We then use these estimates to obtain an upper bound for the *fdr* of a gene.

The unrestricted EM algorithm

We first describe the EM in the full configuration space, and then modify it for the constrained case. That is, the EM is guaranteed to improve the solution and converge. The unrestricted EM formulation is based on repfdr [9, 25], as follows:

The E-step:

$$P(H_{i,\cdot} = h | Z_{i,\cdot}, \pi^{(t)}(h)) = \frac{f(Z_{i,\cdot}|h)\pi^{(t)}(h)}{\sum_{h'} f(Z_{i,\cdot}|h')\pi^{(t)}(h')}$$

The M-step:

$$\pi^{(t+1)}(h) = \frac{1}{n}\sum_i P(H_{i,\cdot} = h|Z_{i,\cdot}, \pi^{(t)}(h)) = \frac{1}{n}\sum_i \frac{f(Z_{i,\cdot}|h)\pi^{(t)}(h)}{\sum_{h'} f(Z_{i,\cdot}|h')\pi^{(t)}(h')}$$

This process guarantees convergence to a local optimum. Our goal is to limit the search space.

Lemma. The EM algorithm above can be used to find a local optimum estimator under the constraint $\forall h \notin \mathcal{H}' \;\; \pi(h) = 0$, for any non-empty configuration set $\mathcal{H}'$.

Proof. Note that during the EM iterations, if at some time point t $\pi^{(t)}(h) = 0$ then $\forall t^* > t$ $\pi^{(t^*)}(h) = 0$. Therefore, setting the starting point of the EM such that $\forall h \notin \mathcal{H}' \; \pi^{(0)}(h) = 0$ satisfies the constraint and ensures convergence.

The restricted EM

In this section we present a restricted version of the EM above, which we call repfdr-UB. To describe it, we need additional notation. Given a configuration vector $h \in \{0, 1\}^m$, let $h[l]$ be the vector containing the first l entries of h. Given a real valued vector v, let $v_{(i)}$ denote the i'th smallest element of v.

Our algorithm is based on the simple observation that if $\pi(h[l]) \leq \epsilon$ then any extension of $h[l]$ cannot exceed ϵ. That is, $\pi(h[l+1]) \leq \epsilon$ regardless of the new value in study $l + 1$. Our algorithm works as follows. The user specifies a limit to the number configurations kept in the memory—n_H. For simplicity we assume that n_H is a power of 2. We first run the unrestricted EM algorithm on the first $log_2(n_H) - 1$ studies. Each subsequent iteration adds a new study. In each iteration l we keep four parameters: (1) $\hat{H}^l$—the set of the top n_H probability configurations, (2) $\hat{\pi}^l$—the vector of their assigned probabilities, (3) $\hat{\xi}^l$—an estimation of $\sum_{h[l]\in\hat{H}^l} \pi^l(h[l])$, and (4) $\hat{\epsilon}^l$—an estimation of the maximal probability among the excluded configurations.

Initially, $l = log_2(n_H) - 1$ and $\hat{H}^l$ contains all possible configurations of the first l studies. In addition $\hat{\xi}^l = 1$, and $\hat{\epsilon}^l = 0$. In iteration $l + 1$ we run the restricted EM algorithm on all possible extensions of $\hat{H}^l$. That is, the input configuration set for the EM is a result of adding either 1 or 0 at the $l + 1$ position of each configuration in $\hat{H}^l$. The EM run produces initial estimations for our parameters, on which the following ordered updates are applied:

$$\hat{\pi}^{l+1} \quad = \quad \hat{\xi}^l \hat{\pi}^{l+1} \tag{1}$$

$$\hat{H}^{l+1} \quad = \quad \{h[l+1]; \hat{\pi}^{l+1}(h[l+1]) \geq \hat{\pi}^{l+1}_{(n_H/2)}\} \tag{2}$$

$$\hat{\xi}^{l+1} \quad = \quad \sum_{h[l+1]\in\hat{H}^{l+1}} \pi^{l+1}(h[l+1]) \tag{3}$$

$$\hat{\epsilon}^{l+1} \quad = \quad \max\,(\hat{\epsilon}^l, \max_{h[l+1]\notin\hat{H}^{l+1}} (\hat{\pi}^{l+1}(h[l+1]))) \tag{4}$$

Note that in step Eq (2) above we keep the top $n_H/2$ configurations in $\hat{H}^{l+1}$. This set is then used as input to the EM run in the next iteration. We repeat the process above until $l = m$. The output of the algorithm is $\hat{H}^m, \hat{\pi}^m, \hat{\xi}^m, \hat{\epsilon}^m$.

A fast algorithm for an upper bound for the fdr. We now show that the restricted algorithm provides an upper bound for the fdr_k of a gene in running time $O(m(k + n_H))$ for each gene.

Theorem. Given the prior probability $\pi(h)$ of each configuration h in $\mathcal{H}' \subseteq \mathcal{H}$, and an upper bound ϵ for the probability of all excluded configurations, the following inequality holds:

$$fdr_k(Z_{i,\cdot}) \leq \frac{\sum_{h:|h|<k \wedge h\in\mathcal{H}'} P(Z_{i,\cdot}|h)(\pi(h)-\epsilon) + \epsilon \sum_{h:|h|<k} P(Z_{i,\cdot}|h)}{\sum_{h\in\mathcal{H}'} P(Z_{i,\cdot}|h)\pi(h)}$$

Proof. Given that for each $h \notin \mathcal{H}'$ $\pi(h) \leq \epsilon$, we get:

$$fdr_k(Z_{i,\cdot}) = \sum_{h:|h|<k \wedge h\in\mathcal{H}'} P(h|Z_{i,\cdot}) + \sum_{h:|h|<k \wedge h\notin\mathcal{H}'} P(h|Z_{i,\cdot}) \leq$$

$$\sum_{h:|h|<k \wedge h\in\mathcal{H}'} P(h|Z_{i,\cdot}) + \epsilon \left(\sum_{h:|h|<k \wedge h\notin\mathcal{H}'} \frac{P(Z_{i,\cdot}|h)}{P(Z_{i,\cdot})} \right)$$

Thus:

$$fdr_k(Z_{i,\cdot}) \leq \sum_{h:|h|<k \wedge h\in\mathcal{H}'} \frac{P(Z_{i,\cdot}|h)\pi(h)}{P(Z_{i,\cdot})} + \epsilon \left(\sum_{h:|h|<k \wedge H\notin\mathcal{H}'} \frac{P(Z_{i,\cdot}|h)}{P(Z_{i,\cdot})} \right) =$$

$$\sum_{h:|h|<k \wedge h\in\mathcal{H}'} \frac{P(Z_{i,\cdot}|h)(\pi(h)-\epsilon)}{P(Z_{i,\cdot})} + \epsilon \left(\sum_{h:|h|<k} \frac{P(Z_{i,\cdot}h)}{P(Z_{i,\cdot})} \right)$$

Finally, since $P(Z_{i,\cdot}) = \sum_h P(Z_{i,\cdot}|h)\pi(h) \geq \sum_{h\in\mathcal{H}'} P(Z_{i,\cdot}|h)\pi(h)$, we obtain:

$$fdr_k(Z_{i,\cdot}) \leq \frac{\sum_{h:|h|<k \wedge h\in\mathcal{H}'} P(Z_{i,\cdot}|h)(\pi(h)-\epsilon) + \epsilon \sum_{h:|h|<k} P(Z_{i,\cdot}|h)}{\sum_{h\in\mathcal{H}'} P(Z_{i,\cdot}|h)\pi(h)}$$

Corollary. The restricted algorithm provides an upper bound for the fdr_k of a gene in running time $O(m(k + n_H))$ for each gene.

Proof. The final term in the proof above can be calculated in $O(m(n_H + k))$. First, the terms $\sum_{h:|h|<k \wedge h\in\mathcal{H}'} P(Z_{i,\cdot}|h)(\pi(h)-\epsilon)$ and $\sum_{h\in\mathcal{H}'} P(Z_{i,\cdot}|h)\pi(h)$ are calculated directly using the output of our EM-like algorithm in $O(mn_H)$. Then, $\epsilon\Sigma_{h:|h|<k} P(Z_{i,\cdot}|h)$ can be calculated using dynamic programming in a similar fashion to our algorithm for calculating fdr_k under independence assumption. Thus, the total running time for all genes, given the output of the EM, is $O(mn(n_H + k))$.

Replicability across independent study clusters

In this section we apply the ideas from the previous sections to obtain an algorithm for calculating the *fdr* under the assumption that the studies originate from independent clusters. We call this algorithm SCREEN (Scalable Cluster-based REplicability ENhancement, see S1 Text for an overview of the method). Briefly, our algorithm has three stages. First, we use the EM-process on each study pair to create a network of study correlations. We then cluster the network to obtain a set of study clusters that are likely to be independent, see Materials and methods for the full description of this step. Second, we run the EM on each cluster separately. Finally, we merge the results from the different clusters using dynamic programming. Note that this algorithm is a heuristic as it uses EM within each cluster.

A fast algorithm for combining study clusters. Assume for now that we are given a clustering of the studies into M independent clusters $C_1, \cdots, C_M$. Thus:

$$\pi(h) = \prod_{j=1}^{M} P(h_{C_j})$$

where h_{C_j} denotes the subvector of h confined to the studies in the cluster C_j. Let Z_{i,C_j} be the z-scores of gene i in the studies of cluster C_j, then fdr_k has the following form:

$$\sum_{h:|h|<k} \prod_{j=1}^{M} \frac{P(Z_{i,C_j}|h_{C_j})P(h_{C_j})}{P(Z_{i,C_j})}$$

This form is a generalization of the formulation under indepepndence assumption, which implies that the dynamic programming approach can be used to merge data across clusters. We now describe the full method.

We apply our EM approach to each cluster separately, and calculate the probability that gene i has exactly k^* non-null realizations in cluster C_j:

$$V_{C_j}[i, k^*] = \sum_{h_{C_j}:|h_{C_j}|=k^*-1} P(h_{C_j}|Z_{i,C_j})$$

Let $V[i, j, k^*]$ be the probability that gene i has $k^* - 1$ non-null realizations over clusters $1, \cdots, j$. Then:

$$\begin{aligned} V[i,1,k^*] &= \begin{cases} V_{C_1}[i,k^*] & \text{if } k^* < |C_1| \\ 0 & \text{otherwise} \end{cases} \\ V[i,j,k^*] &= \sum_{k'=1}^{min(k^*,|C_j|)} V_{C_j}[i,k'] V[i,j-1,k^*-k'] \\ fdr_k(Z_{i,\cdot}) &= \sum_{k^*=1}^{k-1} V[i,m,k^*] \end{aligned}$$

The table V above has $M \times k$ entries for each gene i, and the update rule takes $O(k)$. Thus, given the EM results in each cluster, the running time of this algorithm is $O(k^2M)$ for each gene.

Experimental results

Simulations. Compared methods: We evaluated six different approaches for replicability. (1) Fisher: Fisher's meta-analysis for each gene with BH correction. (2) BH-count: the number of q-values ≤ 0.1 for each gene after applying BH in each study. (3) Exp-count: an estimator for the expected number of non-nulls. See Materials and methods for details on 1-3; (4) Our algorithm for calculating fdr_k exactly under independence assumption, which we call SCREEN-ind, (5) repfdr-UB, which computes an upper bound for fdr_k in order to handle many possibly dependent studies, and (6) SCREEN, which evaluates replicability across study clusters. For 3-6, we tested two methods for two-groups estimation within studies: locfdr [22] and estimation based on mixture of Gaussians, which we call normix, see Materials and methods for details.

Simulation overview: We simulated scenarios of independent and dependent studies as explained later. In each of the scenarios below we started by creating a pair of matrices, P, and H^P. The number of genes n was 5000 and the number of studies m varied (in all scenarios

$m \geq 20$). P is a matrix of p-values, and $H^P_{i,j} \in \{0, 1\}$ denotes whether the p-value of gene i in study j is from the null group ($H^P_{i,j} = 0$) or the non-null group ($H^P_{i,j} = 1$). We first generated H^P, and then determined P given the gene configurations in H^P as follows. For each i, j where $H^P_{i,j} = 0$ the value $P_{i,j}$ was randomly selected from a uniform distribution. For each cell i, j for which $H^P_{i,j} = 1$ the p-value was drawn with probability 0.5 from $\beta(1, x)$ (i.e., low p-values), and otherwise from $\beta(x, 1)$ (i.e., high p-values). We tested $x = 10$, 100, and 1000, corresponding to distributions with 0.1, 0.01, and 0.001 mean p-values, respectively. In addition to the non-null distributions we also varied other parameters in the creation of H^P: the number of non-nulls, and the correlation among the studies (i.e., the columns of H^P).

Performance evaluation: We tested the ability of the methods to detect genes that are non-null in two or more studies. Here, the unknown parameter of a gene was the number k of 1 values in its row in H^P. For each k between 2 and 5 we ran all algorithms above. For repfdr-UB we set the number of configurations n_H to 512. In methods that are based on calculating the local fdr we used a threshold of 0.2 for selecting the genes. For methods that are based on counting we used k as a threshold. For Fisher's meta-analysis we used $q \leq 0.1$ as a threshold. For every k we compared the genes identified by each algorithm to the set of genes with at least k non-null realizations (given in H^P). In each of the cases below we calculated the empirical false discovery proportion (FDP; i.e., the proportion of erroneously declared non-nulls). For methods that had $FDP < 0.2$ across all tested k values we also calculated the Jaccard score in order to quantify the overall agreement between the output and the known solution. Thus, only methods that consistently performed well in terms of FDP are evaluated in their detection power.

Scenario 1: Independent studies. Here, a random set of 300 genes were selected to be non-nulls independently in each study. In addition, we selected 50 genes to have non-null $\beta(1, x)$ p-values in 5 additional studies. These genes were added to ensure presence of a large set of replicable genes with high k values (such a set could arise e.g., from up regulation of a pathway).

Fig 1A shows the results for $x = 100$ using normix. Fisher's method and repfdr-UB had high FDP for $k \geq 3$. Exp-count had better Jaccard than SCREEN and SCREEN-ind, but at the expense of a higher FDP (which reached almost 0.15). For all tested x values the results using locfdr and normix were very similar, and SCREEN was very similar to SCREEN-ind as it tended to correctly cluster each study separately (see Fig 1 and S1–S3 Figs). For $x = 1000$ (S1 Fig) SCREEN and SCREEN-ind had the best Jaccard for almost every k and $FDP \leq 0.1$. In contrast, Fisher's method had very high FDP (≥ 0.25), again illustrating why standard meta-analysis is not suitable for our goal. For $x = 10$ the FDP scores of repfdr-UB were high (e.g., ≥ 0.25) for each k as well as the FDP scores of *SCREEN* for $k = 2$ using the normix method. All other FDP scores were very close to zero, and when the FDP values of SCREEN were high, very few genes where reported(≤ 4), see S3 Fig.

Scenario 2: Dependent studies. We simulated data with $M = 4$ clusters of ten studies each. Here, the matrix H^P was generated by first creating an auxiliary matrix A of the same dimensions as H^P. The rows of A were drawn independently from a multivariate normal distribution $\mathcal{N}(0, \Sigma_M)$, where Σ_M specified a correlation structure of M independent study clusters. Within each cluster the correlation was $r = 0.8$, or $r = 0.4$. H^P was created from A by setting a threshold such that the expected number of non-nulls in each study was 300. That is, $H^P_{i,j} = 1$ if and only if $A_{i,j} \geq \tau^j$, where τ^j is the 0.94-th quantile of the normal distribution of column j of A. Given H^P, P was created as in Scenario 1 with $x = 100$.

The results using normix are summarized in Fig 1B and 1C. In terms of FDP, all algorithms except for Fisher's method and Exp-count performed well. In terms of the Jaccard score SCREEN achieved top performance for $k \geq 4$, and SCREEN-ind achieved top performance for

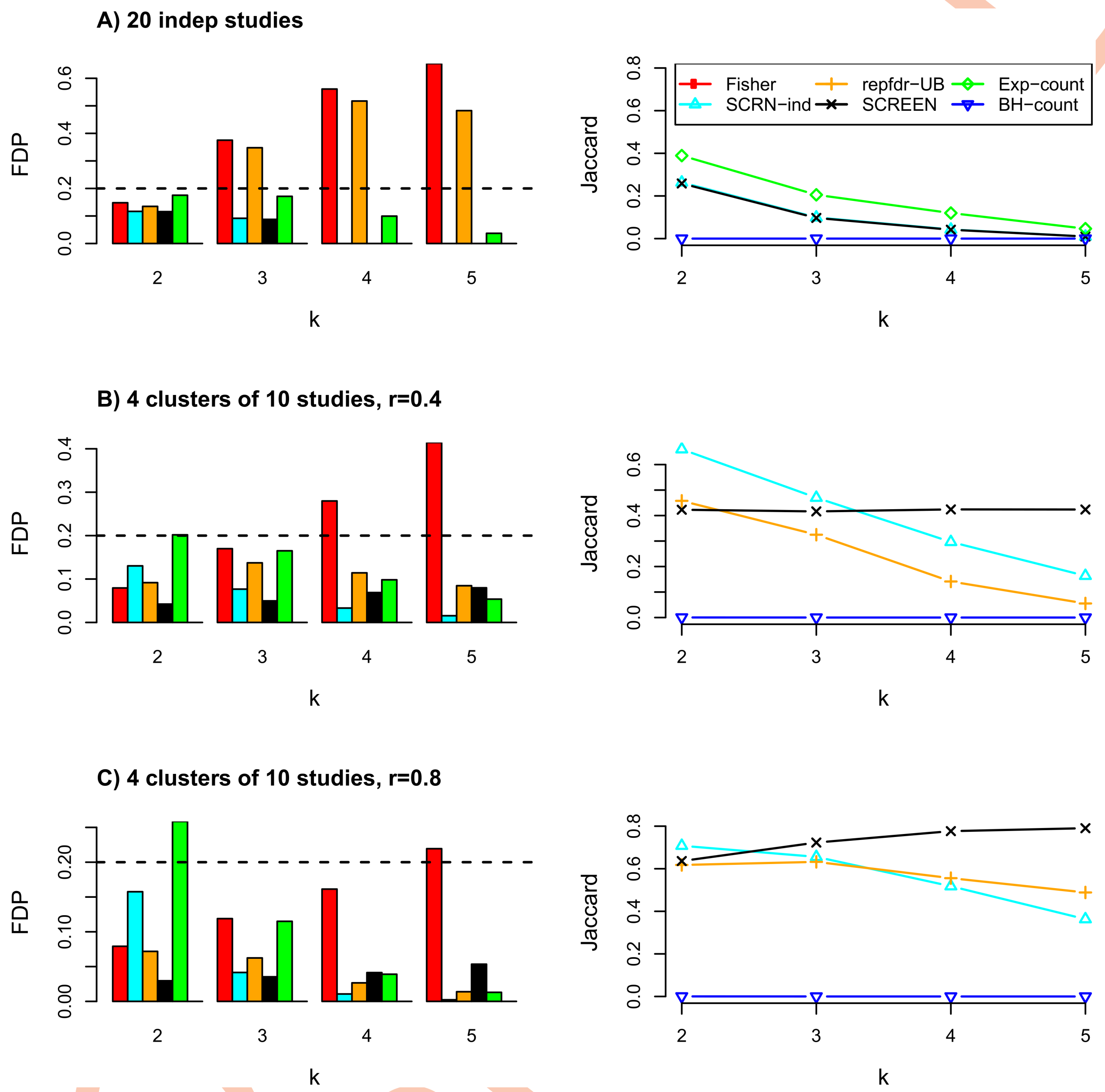


Fig 1. Simulation results: Independent and dependent studies. A) 20 independent studies. B) Four clusters of 10 studies each with a low correlation within each cluster. C) Four clusters of 10 studies each with a high correlation within each cluster. The non-null distribution in each case was Beta(1,100). Two-groups estimation was done using normix. The left column shows the empirical FDP (values above 0.4 are not shown). The right column shows the Jaccard scores only for methods that had consistently low FDP values (< 0.2) for all *k* values. BH-count, SCREEN and SCREEN-ind (SCRN-ind for short) are the only methods that had $FDP \leq 0.2$ in all cases. BH-count had very low FDP but also very low Jaccard scores. Except for $k = 2$, SCREEN had similar or better performance compared to SCREEN-ind.

https://doi.org/10.1371/journal.pcbi.1005700.g001

$k \leq 3$. As in the previous scenario, normix and locfdr gave similar results, see S2 Fig. Fig 2A shows the *fdr* values of SCREEN compared to the real fdr values for different *k* values ($r = 0.8$). For $k = 4$ show that our estimations are highly correlated with the real values. Fig 2B shows that the estimated fdr values decrease with the real number of non-nulls. In addition, most of SCREEN's errors are made for genes with 3 or 4 non-null realizations, and the real fdr values

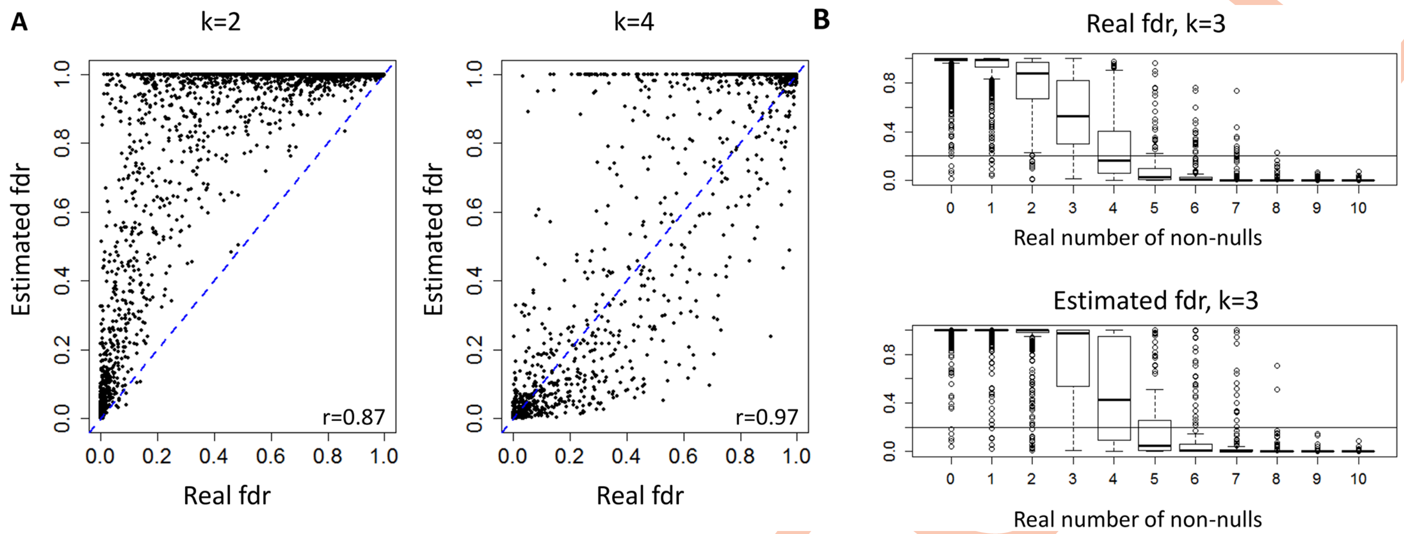


Fig 2. Simulation of dependent studies (4 clusters of 10 studies each). The figures show SCREEN's estimations vs. real fdr values for different *k* values. The number of study clusters is 4 and the correlation within the clusters is set using $r = 0.8$. The non-null distribution within each study is *Beta* (1,100). A) Real vs. estimated fdr values for $k = 2$ and $k = 4$. Each dot corresponds for a gene. For $k = 2$ the estimated *fdr* values are higher, demonstrating stringent FDR control. For $k = 4$ the real and estimated values are highly correlated. B) fdr distributions as a function of the real number of non-nulls (y-axis is the fdr value). Top: real fdrs, bottom: estimated fdr. Each boxplot represents a different set of genes whose real number of non-nulls is given in the x-axis label (except for 10, which means at least 10 non-nulls).

https://doi.org/10.1371/journal.pcbi.1005700.g002

will not necessarily capture these genes at $fdr_k \leq 0.2$. On the other hand, a greater fdr threshold can be used (e.g., 0.4) to cover additional true negatives at the expense of a few false positives.

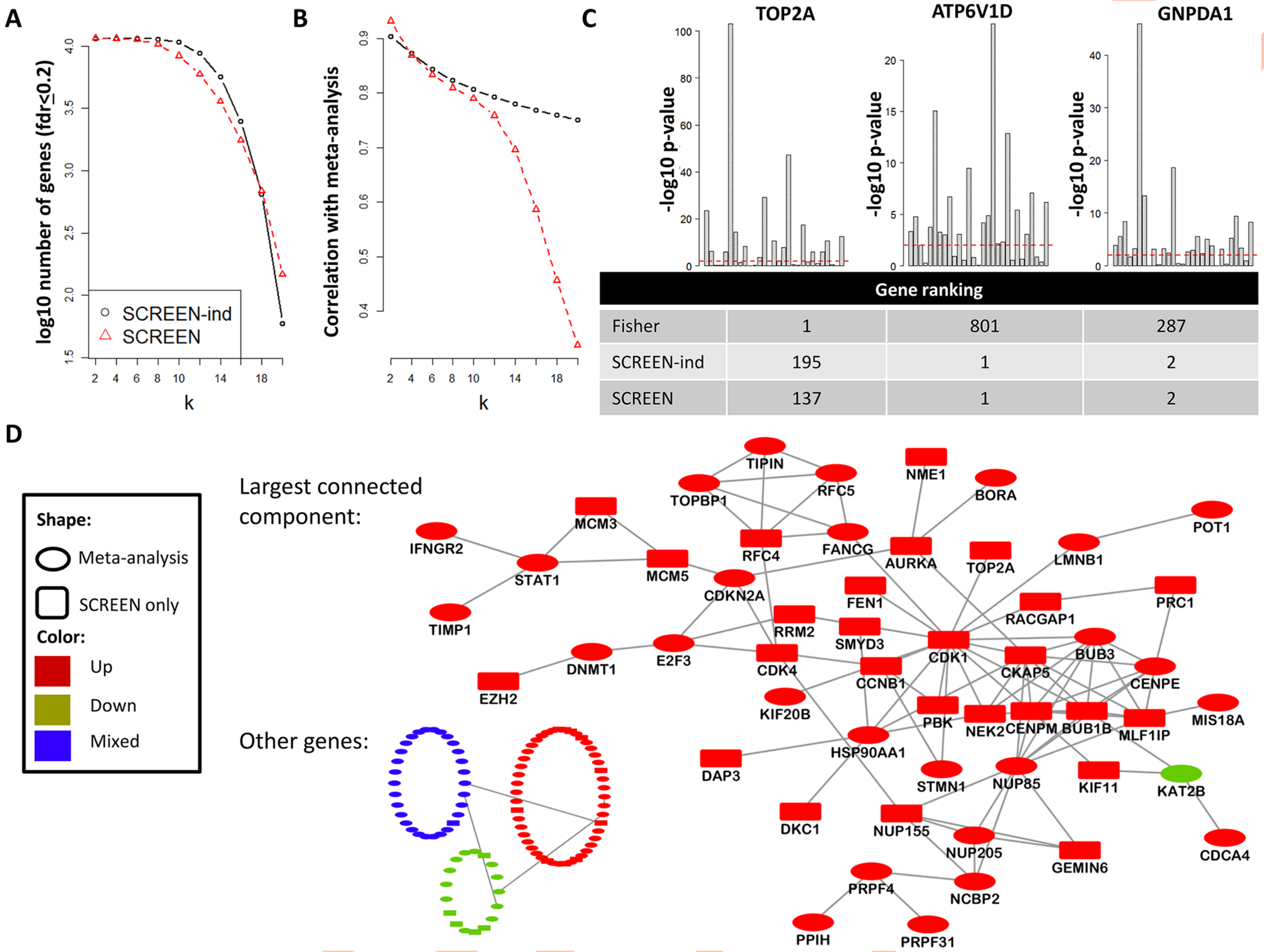
In summary, our simulations show that out of the fdr_k-based methods, SCREEN-ind and SCREEN had consistently low FDP values, while achieving the highest Jaccard scores. SCREEN-ind had a slight advantage in low *k* values, whereas SCREEN was better in higher *k* values.

Scenario 3: Dense effects. In our simulations above the gene effects were sparse. That is, in all scenarios the non-null prior probability was relatively low. When we analyzed real datasets (see below) we observed that while some studies were in line with these classic assumptions, others were not, see S4 and S5 Figs. To cope with these cases of dense effects we performed extensive additional analysis on both simulated and real data, see S1 Text. Our main findings are as follows: (1) locfdr often fails to model these cases, (2) locfdr and normix with empirical null estimation overestimate the null prior probability, and (3) as reported in the previous section, using SCREEN with normix and a fixed theoretical null achieved very high Jaccard scores and low FDP on simulated data. We therefore use the latter approach to analyze the datasets in the subsequent sections.

Analysis of cancer datasets

We analyzed two real datasets. The first, which we call Cancer DEG, is a collection of gene expression studies that compared cancer to non-cancer tissues. The second, called HLA, is from [20], where Shukla et al. tested differential expression between cancer samples with and without somatic mutations in the HLA complex across 11 TCGA cancer subtypes.

The cancer DEG dataset. This dataset contains 29 microarray gene expression studies that compared cancer to non-cancer tissues. It was selected from our previously published compendium [26] by taking all studies that had at least 10 cancer and 10 non-cancer samples (one study was excluded since its gene set was too small). For each dataset genes were assigned



Gene ranking			
Fisher	1	801	287
SCREEN-ind	195	1	2
SCREEN	137	1	2

Fig 3. DEG dataset analysis. A) The number of reported genes at 0.2 *fdr* by SCREEN and SCREEN-ind as a function of *k*. B) The Spearman correlation between gene ranking of SCREEN and SCREEN-ind and of Fisher's meta-analysis as a function of *k*. C) The top ranked genes and their p-values. Top: the p-values of TOP2A, ATP6V1D, and GNPDA1 in each study. Bottom: the rank of these genes according to each of the methods (with k = 20 for SCREEN and SCREEN-ind). ATP6V1D has a very low rank according to Fisher's meta-analysis even though it has consistently low p-values. D) Network analysis of the 147 genes reported by SCREEN with $k = 20$. Nodes are genes, and edges are either protein-protein interactions or known pathway interactions. The largest connected component is shown in detail, and the rest of the genes and their interactions are shown at the bottom left. For each gene we calculated the number of up- and down-regulated t-statistics with a p-value ≤ 0.01. Genes for which the ratio between the up- and down events was ≥ 3 ($\leq 1/3$) were considered consistently up-regulated (down-regulated) in cancer (red and green nodes, 99 and 18 genes, respectively). All other genes were considered as mixed (blue nodes, 30 genes). Oval nodes represent genes ranked among the top 200 genes according to Fisher's meta-analysis (note that even at 10^{-5} Bonferroni correction, more than 10,000 genes were selected in the meta-analysis, We therefore compared to the topmost genes, choosing the number 200 arbitrarily). Rectangular nodes are genes detected only by SCREEN.

https://doi.org/10.1371/journal.pcbi.1005700.g003

p-values for distinguishing between the cancer and non-cancer classes using the NCBI GEO2R web tool [27], where the p-values were calculated using a two-tailed t-test for differential expression. The resulting p-value matrix had 11540 rows (genes) and 29 columns. S6 Fig shows the estimated pairwise correlations between studies. SCREEN identified eight clusters: a single large cluster of 19 studies and 7 clusters with one or more studies.

Fig 3A shows the number of selected genes identified by SCREEN and SCREEN-ind (at *fdr* 0.2), as a function of *k*, the minimum number of studies on which a gene must be detected.

For $k < 10$ the majority of the genes had low *fdr*, suggesting that most genes were differentially expressed in k or more studies. For $k \leq 17$, SCREEN-ind reported more genes than SCREEN. However, SCREEN reported many more genes for higher k values. For example, for $k = 20$ SCREEN-ind detected 59 genes, whereas SCREEN detected 147. In addition, for each k we compared the output of each algorithm to Fisher's meta-analysis. Here, we used Spearman correlation to compare the gene ranking obtained by the methods. Fig 3B shows the results as a function of k. The correlation for low k is near perfect (close to 1 for $k = 2$) and decreases with k. Fig 3C depicts three examples of genes with different ranks: TOP2A, ATP6V1D, and GNPDA1. For each gene the plot shows its $-log_{10}$ p-value in each study, as well as the rank of the gene according to each of the methods. TOP2A was the top ranked gene in Fisher's meta-anlaysis, but had much lower ranks in SCREEN-ind and SCREEN. ATP6V1D and GNPDA1, which were the top two genes of SCREEN and SCREEN-ind, respectively, had much lower ranks in Fisher's meta-analysis. A comparison of the p-value patterns shows that ATP6V1D and GNPDA1 acheived higher rankings even though TOP2A had more extremely low p-values (e.g., $<10^{-20}$). These examples show that SCREEN highlighted genes that were differential consistently across many studies, whereas meta-analysis (as expected) was more sensitive to extremely low p-values. Moreover, the rank-based comparison to Fisher's method shows that selection of genes in such datasets is not only a question of adjusting a threshold, as the gene ranks are very different.

Our discoveries above were based on genes with consistently low p-values in cancer studies. However, the analysis did not use the direction of the differential expression. To shed more light on the directionality we analyzed the 147 genes detected by SCREEN for $k = 20$. For each gene, we compared the times the reported t-statistic was negative and positive (corresponding to down- and up-regulation, respectively). Genes for which the number of negative values was at least thrice (respectively, at most one third of) the number of positive values were denoted as down-regulated (up-regulated). In total, there were 18 down-regulated and 99 up-regulated genes. The remaining 30 genes were denoted as mixed (see S8B Fig).

Using a network of protein-protein interactions and pathway interactions we plotted the subgraph induced by our 147 genes, see Fig 3D. Interestingly, the largest connected component was composed of 52 up-regulated genes and only a single down-regulated gene (KAT2B). Also, 88 out of 94 edges in the network connected up-regulated genes. Many of the genes in this "active" module were not ranked among the top 200 meta-analysis genes obtained using Fisher's method, including genes that are well known to play major roles in cancer formation and progression. For example. CDK1 is a master regulator of cell division, and CKAP5 is important for cytoplasmic microtubule elongation and is known to be over-expressed in colonic and hepatic tumors [28–30]. Other important detected genes that were not among the top 200 meta-analysis genes were CDK1, MCM3, and MCM5. The most enriched GO term in the up-regulated gene set was mitotic cell cycle (38 genes, $q < 10^{-28}$) and the most enriched term in the active module was spindle organization (10 genes, $q = 5.7 \cdot 10^{-11}$).

In summary, our results show that SCREEN revealed a large gene set of consistently up-regulated genes in cancer that are highly relevant in function. On the other hand, while the results of the meta-analysis were also informative they are very sensitive to genes that achieve extreme p-values in one or a few studies and thus may down-weigh consistency. SCREEN was instrumental for separating the main up-regulated cancer genes that were consistent across most cancers, from other genes that manifested few tissue-specific strong effects.

The HLA dataset. Shukla et al. [20] used RNA-seq data to perform differential expression analysis between cancer samples with somatic mutations in the HLA complex and samples without such mutations. The analysis was carried across 11 different TCGA studies, each of a different cancer subtype. The p-value matrix, taken from [20], had 18,128 rows (genes) and 11

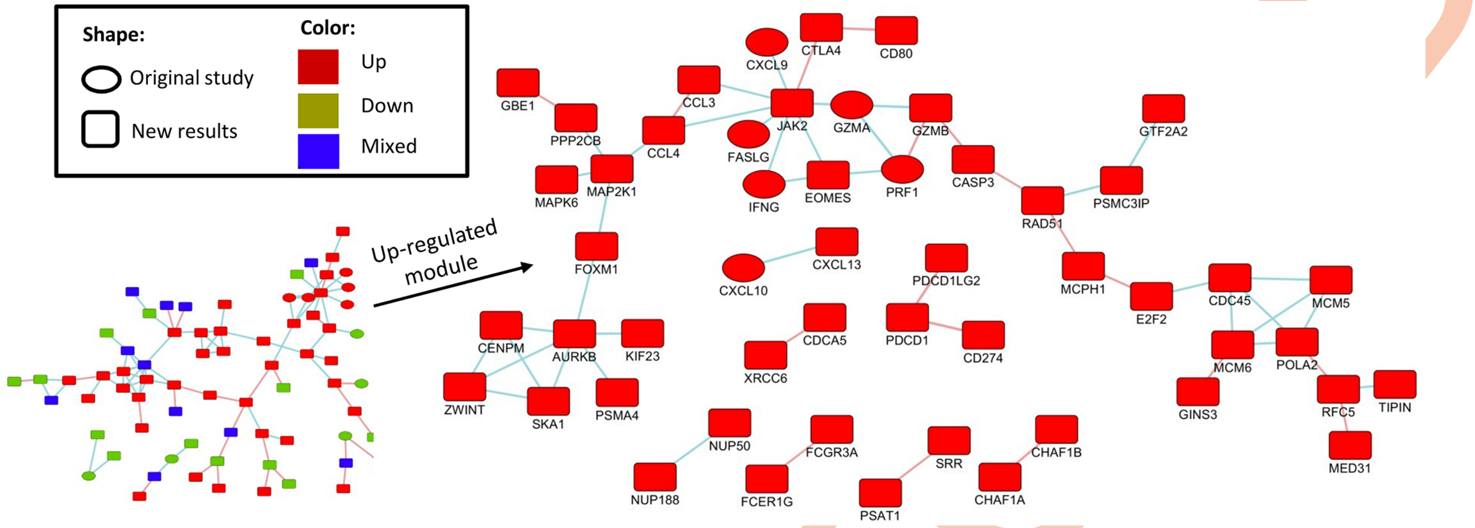


Fig 4. The largest connected component produced by the 405 genes reported by SCREEN in the HLA dataset. Nodes are genes, and edges are either protein-protein interactions or known pathway interactions. Left: all genes in the component, including up-, down-regulated, and mixed genes. Right: the up-regulated genes only. This subnetwork suggests high activity of immune response (which was identified in the original study). Two central genes in the immune response are INFG and JAK2. Our analysis detected both, whereas the original study detected only IFNG. Moreover, the connectivity of the network is established by our newly detected genes.

https://doi.org/10.1371/journal.pcbi.1005700.g004

columns. Similar to the Cancer DEG dataset, the p-values were based on a one-tail test for differential expression: p-values near zero represent up-regulation, and p-values near 1 represent down-regulation. Fisher's method was used to assess the overall significance of a gene, and a total of 119 genes were selected using a p-value cutoff of 10^{-10}.

We applied both SCREEN and SCREEN-ind on these data. Similarly to the DEG dataset, for $k \leq 3$ more than 5000 genes had low *fdr*. However, here the number of genes decreased rapidly with k so that a few hundred were found for $k = 4$, and only a single gene was reported by SCREEN only for $k \geq 5$ (S7 Fig). Unlike the DEG dataset, the overall correlation with Fisher's meta-analysis was lower but the top ranked genes of SCREEN (TNNC2 and IFNG) had also very high ranks in Fisher's method, see S7 Fig.

For $k = 4$ SCREEN reported many more genes than SCREEN-ind (405 vs. 135) genes, and both methods reported more genes than the original analysis (S7D Fig). When performing functional analysis of the 405 genes detected by SCREEN most genes were consistently up-(154) or down-regulated (199), and only 52 genes were mixed. In pathway enrichment analysis our detected gene sets obtained similar results to the original publication and extended them, see S1 Table. Importantly, we recapitulated the main discovery in [20] of up-regulation of multiple immune related processes. Unlike the original publication, our analysis detected enrichment for cancer related pathways in the down-regulated gene sets of SCREEN, including Wnt signaling ($q = 0.02$) and axon guidance ($q = 0.04$).

Fig 4 shows the largest connected component in the network induced by the up-regulated gene set. The network contains a few genes reported in the original study and many newly reported genes, which connect them into a single component. Unlike the DEG dataset, the main connected component contains both up-, down-, and mixed-regulation patterns, but the up-regulated genes form the backbone of the component and are responsible to most of the connecting links. By focusing on the up-regulated genes we detected an active module of genes involved in activation of immune response, anti-tumor activity, and T-cell activation. Many high degree nodes in the module, such as JAK2, were not reported in the original study. In summary, our results provide a comprehensive picture of the molecular response in patients with HLA mutations. SCREEN recapitulated the main findings obtained in the original study and revealed novel ones.

Discussion

We presented here several novel algorithms for detecting replicated associations using an empirical Bayes approach. Our main algorithm, called SCREEN, outperformed other approaches in many scenarios and had consistently low false discovery proportion and high true discovery rates in all simulations.

SCREEN works in three stages. First, it clusters the studies based on their pairwise correlations, which are learned via EM. Second, it performs replicability analysis within each cluster using our restricted EM approach. This method goes beyond previous studies by restricting the possible number of study configurations that are kept in memory. As a result, the method can analyze large study clusters, by computing an upper bound for the *fdr* instead of an exact estimation. Finally, the results of the replicability analyses of the clusters are merged using dynamic programming. For a given k, the output of SCREEN is the fdr_k value for each gene, which can be used to detect genes that are non-null in at least k studies.

We have shown that SCREEN performs well on various simulated scenarios, as well as on real datasets. Specifically, we analyzed two collections of cancer-related gene expression studies. In both cases the discovered gene sets highlighted active gene modules with pertinent functions.Such modules are revealed by the projection of the discovered genes on an interaction network and focusing on well-connected subnetworks. Notably, standard meta-analysis does not reveal many of the genes and generates more fragmented and less coherent subnetworks. For example, in the cancer DEG datasets, some of these genes are central in the network and are known master regulators of cell cycle (e.g., CDK1). In summary, we demonstrated replicability analysis as a standard tool for analyzing a large collection of studies, and provided novel algorithms that are accurate and scalable.

While the current version of SCREEN does not model the direction of the statistic directly (i.e., up- or down-regulation), we addressed this point empirically in our examples and showed that most genes were consistent in their direction. For the sake of functional analysis we required a gene to have the same direction in at least 75% of the studies. This threshold reflects a reasonable selection between the number of reported genes and the consistency requirement (see S8 Fig). Of course, users can change this threshold to require a higher (or lower) consistency in direction in specific applications. Also, note that detection of "mixed sign" genes is an important feature of our analysis: the causality of such genes is questioned as they likely represent downstream effects.

The strategy of SCREEN can be extended to a more complex definition of replicability across study clusters. For example, a researcher may seek genes that are replicable across one or more study clusters, where a gene is replicable in a cluster only if it is non-null in at least some predefined percentage of the studies in that cluster. See S1 Text for a discussion on this topic.

A variety of methods for integration of different studies have been developed in GWAS but their goals are different from SCREEN's. These methods typically assume a standard meta-analysis null hypothesis that the effect size is zero across all studies [31]. As such, they do not address the question of replication directly even if a clustering of the studies is considered [32]. Moreover, some of the Bayesian approaches that were developed for GWAS utilize a subjective prior and do not estimate it form the data [33]. Finally, as we have shown, SCREEN outperforms repfdr, which was originally developed for GWAS data [9].

Our study has some limitations that can be addressed in future research. First, we assumed that genes are independent. This assumption is usually made by state of the art methods, but is often incorrect. In our case, it was used to obtain tractable algorithms (both the dynamic programming and the EM). Second, while our algorithms report fdr_k values of genes, we currently do not estimate their variance. Third, selection of k was done manually on real datasets based on the specific biological question and the number of reported genes (e.g., Fig 3A).

Fourth, our restricted EM approach to analyze study clusters is a heuristic that only guarantees convergence into a local optimum. Thus, while our algorithm has a deterministic starting point for the EM, setting starting points at random may change the obtained upper bounds for the *fdr*. Indeed, when we tried using random starting points, the set of reported genes in the cancer DEG dataset changed (by up to 50 genes, but still leaving over 80 genes from the original results), and no change was observed in the HLA dataset. Another important aspect of the heuristic is the number of allowed gene configurations. As an example, SCREEN with $n_H =$ 10000 configurations on the cancer DEG dataset finds more than 100 additional genes for k = 20, whereas the *fdr* of the genes reported in our original analysis (Fig 4) is kept low, illustrating that the results are consistent (see S9 Fig).

Fifth, while our simple Exp-count approach to estimate the expected number of non-null realizations of a gene performed reasonably well in some simulated scenarios, it is only partially justified theoretically (see S1 Text). Finally, our methods rely on fixed estimates of the two-groups model of each study. Future methods could go a step further and estimate all parameters (i.e., both the study parameters and the gene configuration probabilities) in a single flow.

Materials and methods

Learning a two-groups model in each study

We tried two implementations of two-groups estimation algorithms. The first *locfdr* [22], provides two options to learn the empirical null: maximum likelihood and central matching. By default, we used the maximum likelihood estimator. However, in practice this algorithm might converge to a solution in which $\hat{\pi}_0 > 1$. Whenever this occured, we tried the central matching approach instead. If the new estimator also had $\hat{\pi}_0 > 1$ we used the theoretical null.

The second approach was based on two previous methods: *Znormix* [34] and *fdrtool* [35]. Znormix uses EM to learn a mixture of Gaussians, whereas *fdrtool* assumes that the null distribution is a half normal distribution. Here, we applied an EM approach to the absolute values of the z-scores. We extended these methods by learning a mixture of a half normal with $\sigma \geq 1$ for the null distribution, and a normal distribution with $\mu > 0$ for the non-nulls. We call this approach *normix*.

In practice, we discovered that our EM algorithm is sensitive to high values in the estimation of f_1. In addition, the methods above do not exploit additional information that could be obtained from the two-groups model: an estimation for the power of a study [19]. That is, this is a measure of how separated the two groups are. In our analyses, we took a very stringent approach: in each study we multiply $f_1(z)$ by the estimated power of that study. The effect is a shrinkage in the $f_1(z)$ values that is proportional to the estimated quality of the study.

Clustering the studies

SCREEN relies on a known partition of the studies into clusters. In this section we use an empirical Bayes approach to obtain the clusters. Our analysis has two main parts: learning a network, and clustering.

First, we create a correlation network among the studies. For each study i, let $a_i = P(h_i = 1)$ be the marginal non-null probability in that study. For studies i, j let $a_{i,j} = P(h_i = 1 \wedge h_j = 1)$ be the shared non-null probability of the two studies. We estimate these parameters as follows: a_i and a_j are taken from the two-groups model of each study, and $a_{i,j}$ is estimated by running our EM approach on the data of these two studies. The correlation of the studies is then estimated by:

$$r_{i,j} = \frac{a_{i,j} - a_i a_j}{\sqrt{a_i(1 - a_i)a_j(1 - a_j)}}$$

We obtain a robust estimation of $r_{i,j}$ by taking the mean of 100 bootstrap runs of the procedure above. That is, in each run we reestimate $a_{i,j}$ by running the EM on a bootstrap sample of the genes (n/2 genes out of n, sampled with replacement).

Next, we cluster the network using the infomap algorithm [36]. Here, communities are detected using random walks in the underlying graph. As the input for this algorithm is an unweighted network, we used a threshold of 0.1 for the absolute correlation of study pairs to determine edge presence. This threshold is relatively low for general clustering tasks as it does not guarantee high homogeneity within clusters. However, it guarantees that the clusters discovered by SCREEN will be well-separated. In practice, our clustering approach found the correct clustering of studies in all simulations performed.

Other multi-study analyses

In order to evaluate our *fdr* approaches we compared them to several extant methods for multi-study analysis. Here we outline them briefly.

Fisher's meta-analysis. We used Fisher's meta-analysis to merge the p-values of each gene into a single p-value. We then applied the BH FDR algorithm to account for multiple testing. Note that Fisher's meta-analysis is not meant for replicability analysis and it does not take k as input. Nevertheless, we added it to the comparison due to its use in recent publications (e.g., [20, 37, 38]).

Counting-based BH analysis. Here we run the BH multiple testing correction algorithm in each study separately. For each gene we count the number of q-values lower than some predefined threshold. In this study we used a threshold of 0.1. We call this method BH-count.

Counting-based tdr analysis. In this analysis, for each gene, we use the local true discovery rates (tdr) values obtained from the marginal two-groups model for each study. We then sum over these rates for the gene:

$$\sum_{j=1}^{m}(1 - P(h_{i,j} = 0|Z_{i,j}))$$

This statistic can be interpreted as a biased estimator for the expected number of non-null realizations of gene i. See the S1 Text for more details. We call this method Exp-count.

Enrichment and network analysis

Network analysis and visualization was done in Cytoscape [39]. The GeneMANIA Cytoscape app [40, 41] was used to create the gene networks of the selected gene sets. GO enrichment analysis was performed using Expander [42].

Availability

The datasets and an implementation of SCREEN are available in S1 Data.

Supporting information

S1 Fig. Simulation results: 20 independent studies with $\boldsymbol{\beta}$(1,1000) non-null p-values. A) locfdr, B) normix. The left column shows the empirical FDP. The right column shows the Jaccard scores only for methods that had a consistently low FDP values (< 0.2) in each case and for all k values. These scores are calculated by comparing the output gene set of each method for each k to the set of genes for which the real number of non-nulls was at least k. SCREEN and SCREEN-ind (SCRN-ind) have identical results and achieve the top or almost top Jaccard for $k \leq 4$.
(PDF)

S2 Fig. Simulation results using locfdr and non-null distribution of $\boldsymbol{\beta}$(1,100). A) 20 independent studies. B) Four clusters of 10 studies each with a relatively low correlation within each cluster. C) Four clusters of 10 studies each with a relatively high correlation within each cluster. The left column shows the empirical FDP. The right column shows the Jaccard scores only for methods that had a consistently low FDP values (< 0.2) in each case and for all k values. BH-count, SCREEN and SCREEN-ind (SCRN-ind for short) are the only methods that are shown on the right in all cases. BH-count had very low Jaccard scores. In A) SCREEN-ind had superior results. In B-C) Except for $k = 2$, SCREEN had similar or better performance compared to SCREEN-ind.
(PDF)

S3 Fig. Simulation results: 20 independent studies with $\boldsymbol{\beta}$(1,10) non-null p-values. A) locfdr, B) normix. The left column shows the empirical FDP. The right column shows the Jaccard scores only for methods that had a consistently low FDP values (< 0.2) in each case and for all k values. These scores are calculated by comparing the output gene set of each method for each k to the set of genes for which the real number of non-nulls was at least k. The Jaccard scores here are always very low, and the FDP scores of SCREEN and repfdr-UB might be high. However, when the FDP scores are high, very few genes are reported by SCREEN (≤ 4), whereas repfdr-UB might report ≥ 10 genes. Note that the scale of the Jaccard plots was cleaved to show the very low values.
(PDF)

S4 Fig. P-value histograms of all studies in the cancer DEG dataset.
(PDF)

S5 Fig. P-value histograms of all studies in the HLA dataset.
(PDF)

S6 Fig. Inferred correlation networks among the DEG and HLA studies. Left: inferred correlations. Right: binary correlation. Each point represents correlation with absolute value ≥ 0.1. Study names are colored by their cluster assignment.
(PDF)

S7 Fig. HLA dataset analysis. A) The number of reported genes at 0.2 *fdr* by SCREEN and SCREEN-ind as a function of k. B) The Spearman correlation between gene ranking of SCREEN and SCREEN-ind and of Fisher's meta-analysis as a function of k. C) The top ranked genes and their p-value in each study. Top: the p-values of TNNC2 and IFNG. Bottom: the rank of these genes according to each of the methods (with k = 4 for SCREEN and SCREEN-ind). Both genes are highly ranked by all methods. D) The number of genes reported by each method.
(JPG)

S8 Fig. Differential expression direction analysis of the reported genes. For each reported gene the fraction of datasets with a positive t-statistic is shown (out of the studies with $p < 0.01$). Genes are ordered by the fraction. Red: genes up-regulated in 75% of the studies. Green: genes down-regulated in 75% of the studies. Blue: all other genes. A) HLA dataset, $k = 4$. B) Cancer DEG dataset, $k = 20$.
(PDF)

S9 Fig. Effect of increasing the number of allowed configurations within SCREEN. SCREEN was run with $n_H = 10000$ allowed configurations. The number of detected genes for $k = 20$ more than doubled. While new genes are detected, the genes reported using $n_H = 1024$ only (as in the main text) remain significant (except for 5 genes). The figure shows the distribution of local fdr values for all genes discovered on the cancer DEG dataset with $n_H = 10000$, split into those that were also significant using $n_H = 1024$ and the rest.
(PDF)

S1 Text.
(PDF)

S1 Table. The studies of the cancer DEG dataset.
(PDF)

S1 Data. A zip file with the code of all methods and the matrices of the real datasets.
(ZIP)

Acknowledgments

Part of the work was done while DA and RS were visiting the Simons Institute for the Theory of Computing.

Author Contributions

Conceptualization: David Amar, Daniel Yekutieli.

Formal analysis: David Amar, Ron Shamir, Daniel Yekutieli.

Methodology: David Amar, Ron Shamir, Daniel Yekutieli.

Project administration: Ron Shamir.

Software: David Amar.

Supervision: Ron Shamir.

Validation: David Amar, Daniel Yekutieli.

Visualization: David Amar, Ron Shamir.

Writing – original draft: David Amar, Ron Shamir, Daniel Yekutieli.

Writing – review & editing: David Amar, Ron Shamir, Daniel Yekutieli.

References

1. McNutt M. Reproducibility. Science (New York, NY). 2014; 343(6168):229. https://doi.org/10.1126/science.1250475

2. Camerer CF, Dreber A, Forsell E, Ho Th, Huber J, Kirchler M, et al. Evaluating replicability of laboratory experiments in economics. Science. 2016; 351:1433–1436. https://doi.org/10.1126/science.aaf0918 PMID: 26940865

3. Braver S. Continuously cumulating meta-analysis and replicability. Perspectives on Psychological Science on Psychological Science. 2014; 9(3):333–342. https://doi.org/10.1177/1745691614529796

4. Ioannidis JPA. Why most published research findings are false. PLoS Medicine. 2005; 2:696–701. https://doi.org/10.1371/journal.pmed.0020124

5. Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, et al. Meta-analysis of gene expression profiles in breast cancer: Toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Research: BCR. 2008; 10(4):R65. https://doi.org/10.1186/bcr2124 PMID: 18662380

6. Laas E, Mallon P, Duhoux FP, Hamidouche A, Rouzier R, Reyal F. Low concordance between gene expression signatures in ER positive HER2 negative breast carcinoma could impair their clinical application. PLoS ONE. 2016; 11(2). https://doi.org/10.1371/journal.pone.0148957

7. Verleyen W, Ballouz S, Gillis J. Positive and negative forms of replicability in gene network analysis. Bioinformatics. 2016; 32:1065–1073. https://doi.org/10.1093/bioinformatics/btv734 PMID: 26668004

8. Benjamini Y, Heller R, Yekutieli D. Selective inference in complex research. Philosophical transactions Series A, Mathematical, physical, and engineering sciences. 2009; 367:4255–4271. https://doi.org/10.1098/rsta.2009.0127 PMID: 19805444

9. Heller R, Yekutieli D. Replicability analysis for genome-wide association studies. Ann Appl Stat. 2014; 8 (1):481–498. https://doi.org/10.1214/13-AOAS697

10. Hedges LV, Olkin I. Statistical methods for meta-analysis. Journal of Educational Statistics. 20(1); 1985.

11. Chang LC, Lin HM, Sibille E, Tseng GC. Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. BMC Bioinformatics. 2013; 14(1):368. https://doi.org/10.1186/1471-2105-14-368 PMID: 24359104

12. Li Y, Ghosh D. Meta-analysis based on weighted ordered P-values for genomic data with heterogeneity. BMC Bioinformatics. 2014; 15(1):226. https://doi.org/10.1186/1471-2105-15-226 PMID: 24972803

13. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science (New York, NY). 2007; 316(5829):1336–41. https://doi.org/10.1126/science.1142364

14. Kraft P, Zeggini E, Ioannidis JPa. Replication in Genome-Wide Association Studies. Statistical Science. 2009; 24(4):561–573. https://doi.org/10.1214/09-STS290 PMID: 20454541

15. Benjamini Y, Heller R. Screening for partial conjunction hypotheses. Biometrics. 2008; 64(4):1215–1222. https://doi.org/10.1111/j.1541-0420.2007.00984.x PMID: 18261164

16. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B. 1995; 57(1):289–300.

17. Heller R, Bogomolov M, Benjamini Y. Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111:16262–16267. https://doi.org/10.1073/pnas.1314814111 PMID: 25368172

18. Song C, Tseng GC. Hypothesis setting and order statistic for robust genomic meta-analysis. Annals of Applied Statistics. 2014; 8(2):777–800. https://doi.org/10.1214/13-AOAS683 PMID: 25383132

19. Efron B. Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction. Institute of Mathematical Statistics monographs. Cambridge: Cambridge University Press; 2010.

20. Shukla SA, Rooney MS, Rajasagi M, Tiao G, Dixon PM, Lawrence MS, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nature Biotechnology. 2015; 33:1152–1158. https://doi.org/10.1038/nbt.3344 PMID: 26372948

21. Storey JD. The optimal discovery procedure: A new approach to simultaneous significance testing. Journal of the Royal Statistical Society. Series B: Statistical Methodology. 2007; 69:347–368. https://doi.org/10.1111/j.1467-9868.2007.005592.x

22. Efron B. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association. 2004; 99(465):96–104. https://doi.org/10.1198/016214504000000089

23. Efron B. Correlation and large-scale simultaneous significance testing. Journal of the American Statistical Association. 2007; 102(477):93–103. https://doi.org/10.1198/016214506000001211

24. Efron B. Empirical Bayes estimates for large-scale prediction problems. Journal of the American Statistical Association. 2009; 104(487):1015–1028. https://doi.org/10.1198/jasa.2009.tm08523 PMID: 20333278

25. Yekutieli D. repfdr: a tool for replicability analysis for genome-wide association studies. Bioinformatics (Oxford, England). 2014; 30:2971–2972. https://doi.org/10.1093/bioinformatics/btu434

26. Amar D, Hait T, Izraeli S, Shamir R. Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets. Nucleic Acids Research. 2015; 43(16):7779–7789. https://doi.org/10.1093/nar/gkv810 PMID: 26261215

27. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: Archive for functional genomics data sets—Update. Nucleic Acids Research. 2013; 41. https://doi.org/10.1093/nar/gks1193 PMID: 23193258

28. Charrasse S, Mazel M, Taviaux S, Berta P, Chow T, Larroque C. Characterization of the cDNA and pattern of expression of a new gene over-expressed in human hepatomas and colonic tumors. European Journal of Biochemistry / FEBS. 1995; 234:406–13. https://doi.org/10.1111/j.1432-1033.1995.406_b.x

29. Enserink JM, Kolodner RD. An overview of Cdk1-controlled targets and processes. Cell Division. 2010; 5:11. https://doi.org/10.1186/1747-1028-5-11 PMID: 20465793

30. Royle SJ. The role of clathrin in mitotic spindle organisation. Journal of Cell Science. 2012; 125:19–28. https://doi.org/10.1242/jcs.094607 PMID: 22294613

31. Evangelou E, and Ioannidis J P A. Meta-analysis methods for genome-wide association studies and beyond. Nature Reviews Genetics. 2013; 14:379–389. https://doi.org/10.1038/nrg3472 PMID: 23657481

32. Morris AP. Transethnic meta-analysis of genomewide association studies. Genetic Epidemiology. 2011; 35:809–822. https://doi.org/10.1002/gepi.20630 PMID: 22125221

33. Yang J, and Lee SH, Goddard ME, and Visscher PM. GCTA: A tool for genome-wide complex trait analysis. American Journal of Human Genetics. 2011; 88:76–82. https://doi.org/10.1016/j.ajhg.2010.11.011 PMID: 21167468

34. McLachlan GJ, Bean RW, Jones LBT. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. Bioinformatics. 2006; 22(13):1608–1615. https://doi.org/10.1093/bioinformatics/btl148 PMID: 16632494

35. Strimmer K. A unified approach to false discovery rate estimation. BMC Bioinformatics. 2008; 9(1):303. https://doi.org/10.1186/1471-2105-9-303 PMID: 18613966

36. Rosvall M, Bergstrom CT. An information-theoretic framework for resolving community structure in complex networks. PNAS. 2007; 104:7327. https://doi.org/10.1073/pnas.0611034104 PMID: 17452639

37. Schunkert H, König IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nature Genetics. 2011; 43:333–338. https://doi.org/10.1038/ng.784 PMID: 21378990

38. Kaever A, Landesfeind M, Feussner K, Morgenstern B, Feussner I, Meinicke P. Meta-analysis of pathway enrichment: Combining independent and dependent omics data sets. PLoS ONE. 2014; 9. https://doi.org/10.1371/journal.pone.0089297 PMID: 24586671

39. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research. 2003; 13 (11):2498–2504. https://doi.org/10.1101/gr.1239303 PMID: 14597658

40. Montojo J, Zuberi K, Rodriguez H, Kazi F, Wright G, Donaldson SL, et al. GeneMANIA cytoscape plugin: Fast gene function predictions on the desktop. Bioinformatics. 2010; 26:2927–2928. https://doi.org/10.1093/bioinformatics/btq562 PMID: 20926419

41. Vlasblom J, Zuberi K, Rodriguez H, Arnold R, Gagarinova A, Deineko V, et al. Novel function discovery with GeneMANIA: a new integrated resource for gene function prediction in Escherichia coli. Bioinformatics (Oxford, England). 2014; p. 1–5.

42. Ulitsky I, Maron-Katz A, Shavit S, Sagir D, Linhart C, Elkon R, et al. Expander: from expression microarrays to networks and functions. Nature Protocols. 2010; 5:303–322. https://doi.org/10.1038/nprot.2009.230 PMID: 20134430