Novel Statistical Tools for Management of Public Databases Facilitate Community-Wide Replicability and Control of False Discovery

Saharon Rosset* Ehud Aharoni†

Abstract

Issues of publication bias, lack of replicability and false discovery have long plagued the genetics community. Proper utilization of public and shared data resources presents an opportunity to ameliorate these problems. We present an approach to public database management that we term Quality Preserving Database (QPD). It enables perpetual use of the database for testing statistical hypotheses while controlling false discovery and avoiding publication bias on the one hand, and maintaining testing power on the other hand. We demonstrate it on a use case of a replication server for GWAS findings, underlining its practical utility. We argue that a shift to using QPD in managing current and future biological databases will significantly enhance the community's ability to make efficient and statistically sound use of the available data resources.

The prevalence and importance of public databases and shared data resources in the genetics and molecular biology community have been constantly increasing, and this trend is expected to continue and accelerate with the growth in genetic data generation. For example, publicly available GWAS data sources like WTCCC (Wellcome Trust Case Control Consortium, 2010) and dbGaP (Mailman et al., 2007) have had a major contribution in genetic research well beyond the findings in the original studies which collected the data. An important class of data sharing efforts is the formation of scientific consortia for collection and analysis of genetic data, ranging in size from several collaborating scientists to entire research communities (Rich et al., 2007). The coming years are expected to see the formation of multi-purpose public genetic databases of unprecedented size and utility (Couzin-Frankel, 2012).

*Department of Statistics and Operations Research, Tel Aviv University 69978 Israel. Email: saharon@post.tau.ac.il
†IBM Research Laboratory, Haifa, Israel

In parallel to this increasing data availability and data-sharing trend, a growing concern in medicine, genetics and other sciences surrounds the validity and replicability of scientific findings and scientific publications (Yong, 2012). One major issue is multiple comparisons and false discoveries. Fundamentally, when numerous hypotheses are being tested in the same study (like 500,000 tests for SNP association in a typical GWAS) or across many different studies, lack of proper and careful correction for multiple comparisons practically guarantees false discoveries, publication bias and lack of replicability (Ioannidis, 2005).

In our view, shared or public data resources represent a tremendous opportunity to change the culture of genetic research, and ensure statistical validity of scientific results, with minimal efforts. The key idea is one we term Quality Preserving Database (QPD), whose main premise is that we want to make our shared data resource available for unlimited and perpetual use for testing statistical hypotheses while remaining useful (maintaining power) on the one hand, and ensuring statistical validity of findings across all tests performed on the other hand. Within a QPD framework, the ability of future users to perform tests and make valid discoveries is not negatively impacted by tests being performed now.

From first principles, a QPD is impossible without incurring some cost. In our scheme, usage of the QPD for testing incurs a cost in the form of additional samples (or equivalently, cost of acquiring samples). It is this resulting gradual growth of the number of samples which enables perpetual use while provably controlling false discovery and maintaining testing power (Fig. 1, see Aharoni et al. (2011) and Methods for technical details). Summed up in one sentence, a QPD is a database with a management layer that fairly assigns costs in the form of additional data samples (or cost of acquiring them) for each test executed, and controls some measure of false discovery across all tests performed.

A key radical feature in the QPD approach is that the scientists performing the tests are no longer concerned with p values and are not directly responsible for controlling type-I errors — these aspects are now the responsibility of the database manager. Rather, they are focused on what they wish or expect to find: effect size and power. According to their specifications of these quantities, the database manager is able to assign a cost to the test. The validity of the test (and hence the calculation of p value and determination of thresholds) are the responsibility of the database manager and consequently, the cost of the test depends on its "difficulty": high power for low effect size would be much more expensive. Furthermore, a critical point in our view is that the database manager does not need to examine the scientific merit of the hypotheses being tested, because the QPD scheme guarantees future users will not be hampered by tests being performed now. As long as the scientist is willing to shoulder the cost of the test (in samples or equivalent payment), there is no harm in performing it. Consequently, a QPD management scheme does not require a scientific

review of the tests being performed.

A relatively simple use case for QPD is as a replication server for GWAS findings. Common practice in the genetics community calls for replication of any major scientific finding on a second independent dataset before it is fully accepted (Lander and Kruglyak, 1995). A community of researchers of a specific disease could thus establish a shared replication server, managed as a QPD. In fact, it may be pertinent for central research authorities like the NIH to take responsibility for setting up such a replication server. Because it is managed as a QPD, we can be confident of all replications successfully performed on it, while their cost is kept minimal. As an illustration, we can compare three approaches for replicating GWAS findings:

1. Collect brand new data and attempt the replication on it, publish if successful (that is, if the result is significant)

2. Obtain publicly available data, attempt the replication on it, publish if successful

3. Use a replication QPD, publish if successful

For simplicity of statistical exposition we assume here that replication involves a standard test of a normal mean, although realistic tests used in replication (such as likelihood ratio tests for case-control data) would behave similarly. For such a generic testing scenario (normal test, effect size 0.1, power 0.95), option 1 requires 1082 samples for a test at level 0.05. However, it is subject to publication bias and lacks false discovery control across attempted replications by different researchers. Hence this approach to replication is both expensive and unreliable. Option 2 requires no data collection, but in addition to possible data quality and heterogeneity concerns, is subject to severe publication bias, assuming many different researchers are attempting replications on this same data. Option 3 typically costs less than two samples or equivalent payment (Fig. 2, see Methods for calculations description), and since it controls false discovery across all tests performed on the QPD, the researcher is protected against publication bias even if only successful replications are published. Furthermore, if the community or research authority takes it upon itself to supply a steady stream of additional samples into the replication server, it could negate the need to charge a payment for use, while maintaining false discovery control.

The typical required payment for QPD use is between zero and three samples per test performed (Fig. 2). This "typical" cost depends on several factors, in particular the measure of false discovery being controlled. Our first published QPD schema controls the family-wise error rate (FWER), the probability of making even one false discovery (Aharoni et al., 2011). This very conservative measure results in relatively slowly decreasing costs. It is widely appreciated in the genetics community that controlling less conservative

measures of false discovery than FWER can lead to more valid discoveries. In particular, the false discovery rate (FDR) is a popular measure. In FDR, the quantity being controlled is the expected portion of discoveries that are not valid (compared to the probability of making even one false discovery in FWER). Unfortunately, there are no known schemes for controlling FDR in a sequential scenario like QPD. However, a slightly modified measure called mFDR can be controlled sequentially under some assumptions (Foster and Stine, 2008), and implemented within a QPD framework (Aharoni and Rosset, 2013). The intriguing consequence is that when controlling mFDR, the cost of perpetual use of a QPD can become zero indefinitely, as long as some "good" tests (that is, truly novel tests with false nulls and good power) are regularly performed on it (Fig. 2).

The QPD approach, while radical, integrates naturally with some emerging trends in public database management, emphasizing privacy and cost considerations. It requires a management layer, which controls access to the database, and a shift in how tests are defined and performed on such a database. While researchers are currently engrained in the "p-value culture", we believe the quantities the QPD requires them to specify — effect size and power — are actually more natural for scientists than p values, as they refer to the discoveries they hope to make (and are already implicitly involved in deciding which tests to perform). Furthermore, if the researcher is unable to specify a desired effect size and power combination, the database manager can specify different combinations and their prices. In the replication server example, the researcher usually has an idea of the effect size hypothesized from the initial study, and choosing a power for the test can then be thought of as a cost-benefit tradeoff — the more a researcher can pay, the higher the power that can be guaranteed.

A critical aspect of the QPD scheme, which requires a cautionary note, is the assumption that additional relevant samples can be provided by the QPD users, or acquired by the database manager. All samples, including those currently in the database and those added to account for usage, are assumed to be of equal quality and representative of a common distribution. Thus, aspects like data heterogeneity, population structure and limited availability of samples (say, cases from a rare disease) should be carefully considered when designing a QPD.

In our view, the benefits of using QPD — perpetual statistical validity and efficient use of data — far outweigh the cultural and administrative challenges it presents to the community.

Methods

This section is based on our published papers (Aharoni et al., 2011; Aharoni and Rosset, 2013) defining QPD and its implementations controlling FWER and mFDR. We cite the relevant results from both papers, and refer the reader to the papers themselves for more details and proofs.

Definitions

The usual setting of false discovery analysis assumes m null hypotheses, H_1, H_2,...,H_m. Let Θ_j be the parameter space assumed by the j'th test, and the null hypothesis H_j is defined as a subset $H_j \subset \Theta_j$. A random variable $R_j \in \{0,1\}$ indicates whether H_j was rejected, and $R = \sum_{j=1}^{j=m} R_j$ counts the total number of rejections. Similarly, $V_j \in \{0,1\}$ indicates the case that H_j is true and is rejected (i.e., type-I error), and $V = \sum_{j=1}^{j=m} V_j$ tracks the total number of type-I errors.

The two most commonly used measures of false discovery are the Family-Wise Error Rate (FWER) and False Discovery Rate (FDR), defined as follows.

$$\begin{aligned} FWER &\equiv P(V > 0) \\ FDR &\equiv E(\tfrac{V}{R}|R > 0)P(R > 0). \end{aligned} \tag{1}$$

FWER is the probability of making one or more false discoveries, while FDR is the expected percentage of false discoveries among the total number of discoveries.

In the search for a sequential procedure for controlling FDR, a modified measure has been employed, $mFDR_\eta$ (Foster and Stine, 2008):

$$mFDR_\eta \equiv \frac{E(V)}{E(R) + \eta}, \tag{2}$$

where $\eta > 0$ is some constant, typically chosen to be $\eta = 1 - \alpha$.

Control of $mFDR$ requires assumption 1, which can be viewed as a weaker form of requiring independence between tests.

Assumption 1

$$\forall \theta \in H_j : P_\theta(R_j | R_{j-1}, R_{j-2}, \ldots, R_1) \leq \alpha_j \tag{3}$$

$$\forall \theta \notin H_j : P_\theta(R_j | R_{j-1}, R_{j-2}, \ldots, R_1) \leq \rho_j, \tag{4}$$

where θ is the combined true parameter values of all m tests, and $P_\theta(\cdot)$ and $E_\theta(\cdot)$ denote the probability

distribution and expectation when assuming θ. We denote the level of the j'th test by α_j, and ρ_j is the maximal power, defined as follows:

$$\rho_j = sup_{\theta_j \in \Theta_j - H_j} P_{\theta_j}(R_j = 1). \tag{5}$$

The Quality Preserving Database

A QPD (Aharoni et al., 2011) serves a series of test requests. Each request includes the following information: The test statistic, assumptions on the distribution of the data (e.g. that it is normally distributed), the desired effect size and power requirements. The request does *not* contain two details: the required significance level and the number of samples which will be used to calculate the test statistic. These two details are managed by the QPD's manager. However, given all the other request details the significance level becomes a function of the number of samples. We term this function the "level-sample" function, formally defined below.

Definition 2 *Given a test request containing test statistic, data assumptions, and the desired effect size and power requirements, the* Level-sample *function, $L : \mathbb{N} \to \mathbb{R}$, is a function specifying the feasible significance level given the number of samples.*

The level-sample function summarizes the test request for the sake of determining the cost required for executing it. The costs assigned for requests may vary between different requests, and also may vary with time, but the allocation scheme must fulfill two properties, *fairness* and *stability*, defined formally next.

Definition 3 *The costs assigned by the QPD satisfy the* fairness *requirement if for any two requests such that one has level-sample function $L_a(n)$ and the other $L_b(n)$, and $\forall n : L_a(n) < L_b(n)$, then at any particular point in time the first request will be assigned with no higher cost than the second request.*

Definition 4 *The costs assigned by the QPD satisfy the* stability *requirement if for any particular request a there is some constant c_a such that the cost assigned to it will never exceed c_a.*

Now a QPD can be formally defined as follows.

Definition 5 *A Quality Preserving Database (QPD) is a database with a management layer that assigns costs in the form of additional data samples for each test executed. This layer fulfills the following three properties: (a) It can serve an infinite series of requests, (b) It satisfies the fairness and stability requirements, (c) It controls some measure of the overall type-I errors (e.g., $FWER$) at some pre-configured level α.*

In previous publications we have shown multiple ways to fulfil the above definition.

In Aharoni et al. (2011) we have shown a version we term here $QPD\text{-}AS(\alpha, q)$, that uses Alpha Spending to control FWER. Let $W(j)$ be the α-wealth remaining after the j'th test, i.e., $W(j) = W(j-1) - \alpha_j$. $W(0)$ is the initial α-wealth set by the QPD manager, which guarantees $FWER \leq W(0)$. In exchange for a payment of c_j new samples, the j'th user receives an α_j allocation of:

$$\alpha_j = W(j-1)(1 - q^{c_j}), \tag{6}$$

where $0 < q < 1$ is a parameter.

Theorem 6 *(Aharoni et al., 2011) $QPD\text{-}AS(\alpha, q)$ fulfills the three properties of Definition 5, where stability is only guaranteed for requests such that their level-sample function $L : \mathbb{N} \to \mathbb{R}$ satisfies $L(n) \leq bq^n$ for some b. $QPD\text{-}AS(\alpha, q)$ always controls FWER at level α.*

The requirement for level sample exponential decay decay is fulfilled by a wide range of commonly used statistical tests, among them Neyman-Pearson tests, tests about the mean of a normal distribution with known variance, and single-tail, uniformly most powerful tests with one unknown parameter (Aharoni et al., 2011). Simulations show that many more types of tests are applicable in practice, e.g., tests based on an approximately normal distribution such as a t-distribution.

In Aharoni and Rosset (2013) we present two additional variants $QPD\text{-}ASR(\alpha, \eta, q)$ and $QPD\text{-}ASR\text{-}OPT(\alpha, \eta, q)$ that use a novel approach we termed *Alpha Investing with Rewards* to control mFDR.

$QPD\text{-}ASR(\alpha, \eta, q)$ differs from $QPD\text{-}AS(\alpha, q)$ in the way the wealth $W(0)$ is managed. It is initialized with $W(0) = \alpha\eta$, which is typically lower than α (e.g., when $\eta = 1 - \alpha$), but when a rejection occurs it gets an additional wealth bonus amount of α.

In $QPD\text{-}ASR\text{-}OPT(\alpha, \eta, q)$, we split the α-wealth $W(j)$ into two pools of wealth, $W(j) = A(j) + B(j)$. The first pool, $A(j)$, is the same α-wealth that a $QPD\text{-}AS(\alpha\eta, q)$ would have. The bonus amounts of α obtained by rejecting hypotheses are added to the second pool, $B(j)$. The α_j allocation is combined of two allocations from the two pools. The amount withdrawn from the $A(j)$ pool is the same as in the $QPD\text{-}AS$, i.e., $A(j-1)(1 - q^{c_j})$. The $B(j)$ pool, on the other hand, is managed more openhandedly, knowing it will be refilled with α every time a rejection occurs. At each step we estimate the probability of this happening based on prior history $p(j) = \frac{1}{j-1}\sum_{i=1}^{j-1} R_j$, and withdraw the following amount for the j'th test: $min(p(j)\alpha, B(j-1))$. Albeit somewhat heuristic, this management of the $B(j)$ pool does have the following property. If we may assume a real probability of rejection p, then the expected amount withdrawn at the

j'th step from $B(j)$ is $E(\min(p(j)\alpha, B(j-1))) \leq E(p(j)\alpha) = p\alpha$. The expected reward of the j'th test is also $p\alpha$. Therefore the $B(j)$ pool is kept in a balanced manner. Note that the amounts withdrawn from it are irrespective of c_j, and therefore can be thought of as 'free of charge'. We show in Aharoni and Rosset (2013) that in practical scenarios the costs of tests reduce to zero thanks to this management of the $B(j)$ pool.

The following equations formally define the level allocation of the j'th test in $QPD\text{-}ASR\text{-}OPT(\alpha, \eta, q)$.

$$\begin{aligned} p(j) &= \tfrac{1}{j-1}\textstyle\sum_{i=1}^{j-1} R_i \\ \alpha_j &= A(j-1)(1-q^{c_j}b) + \min(p(j)\alpha, B(j-1)). \end{aligned} \tag{7}$$

The following theorem states both of these variants fulfil the QPD definition as well.

Theorem 7 *(Aharoni and Rosset, 2013) Both $QPD\text{-}ASR\text{-}OPT(\alpha, \eta, q)$ and $QPD\text{-}ASR(\alpha, \eta, q)$ fulfill the three properties of Definition 5, where stability is only guaranteed for requests such that their level-sample function $L : \mathbb{N} \to \mathbb{R}$ satisfies $L(n) \leq bq^n$ for some b, and type-I error control is satisfied by controlling $mFDR$ at level α given assumption 1.*

Example: calculations in Figure 2 of our main text

This figure was created by assuming a series of tests of normal means with variance 1 on a database with a starting size of 2000 samples, where each test is required to guarantee power of 0.95 at effect size 0.1. We first considered a $QPD\text{-}AS(0.05, 0.999)$ scheme, where the cost of each test in the sequence is deterministic and does not depend on the outcome of previous tests or on the prevalence of false null hypotheses in the tested sequence. Explicitly, the $QPD\text{-}AS$ scheme determines the payment of user j by calculating the level that should be allocated to the test:

$$\alpha_j = \Phi(Z_{0.95} - 0.1\sqrt{n_j}),$$

and then the number of samples this allocation requires as payment is the minimal value c_j that gives:

$$\Phi(Z_{0.95} - 0.1\sqrt{n_{j-1} + c_j}) \leq W(j-1)(1 - 0.999^{c_j}).$$

This is the solid line in the figure, and it guarantees FWER control at level 0.05 for the infinite sequence.

We further simulated $QPD\text{-}ASR\text{-}OPT(0.05, 0.95, 0.999)$. Here the frequency of true discoveries matters, and our streams of independent normal tests had probability 0.1 of being a false null (i.e., a potential true

discovery). For the *QPD-ASR-OPT* implementation, the rewards of Alpha Spending with Rewards are accumulated in the $B(j)$ pool. When serving a new request, we first extract free of charge from $B(j)$ at most $p(j)\alpha$, where $p(j)$ is an estimate of the probability of rejection, gradually converging to 0.1. If this is not enough to perform the test at the required power, a cost c_j is computed to extract the remainder from the second pool $A(j) = W(j) - B(j)$, in the same manner as above. The dashed line in Figure 2 is the average cost from averaging 50 such independent simulation streams. Starting from about the 50th test, the use was free in all 50 simulations, as the $B(j)$ pool was consistently non-empty and allowed assignment of sufficient levels to all tests.

References

Aharoni, E., H. Neuvirth, and S. Rosset (2011). The quality preserving database: A computational framework for encouraging collaboration, enhancing power and controlling false discovery. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on 8*(5), 1431–1437.

Aharoni, E. and S. Rosset (2013). Generalized alpha investing: Definitions, optimality results, and application to public databases. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. To appear, accepted 07/13.

Couzin-Frankel, J. (2012, December). U.K. unveils plan to sequence whole genomes of 100,000 patients. Science News Online. http://news.sciencemag.org/biology/2012/12/u.k.-unveils-plan-sequence-whole-genomes-100000-patients.

Foster, D. P. and R. A. Stine (2008, January). Alpha-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(2), 429–444.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine 2*(8), e124.

Lander, E. and L. Kruglyak (1995, November). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature genetics 11*(3), 241–247.

Mailman, M. D., M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, N. Popova, S. Pretel, L. Ziyabari, M. Lee, Y. Shao, Z. Y. Wang, K. Sirotkin, M. Ward, M. Kholodov, K. Zbicz, J. Beck, M. Kimelman, S. Shevelev, D. Preuss, E. Yaschenko, A. Graeff, J. Ostell,

and S. T. Sherry (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nature genetics 39*(10), 1181–1186.

Rich, S. S., P. Concannon, H. Erlich, C. Julier, G. Morahan, J. Nerup, F. Pociot, and J. Todd (2007, October). The type 1 diabetes genetics consortium. *Ann N Y Acad Sci. 1079*, 1–8.

Wellcome Trust Case Control Consortium (2010, April). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature 464*(7289), 713–720.

Yong, E. (2012, May). Replication studies: Bad copy. *Nature 485*(7398), 298–300.

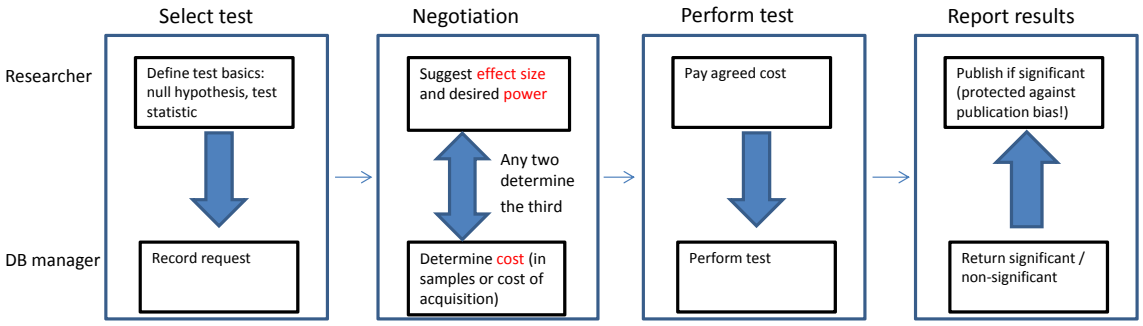

Figure 1: Process of performing a single test on a public database managed as a QPD. If the DB manager uses a QPD scheme in determining cost and performing the test, it guarantees overall control of false discovery on all tests performed on the database.

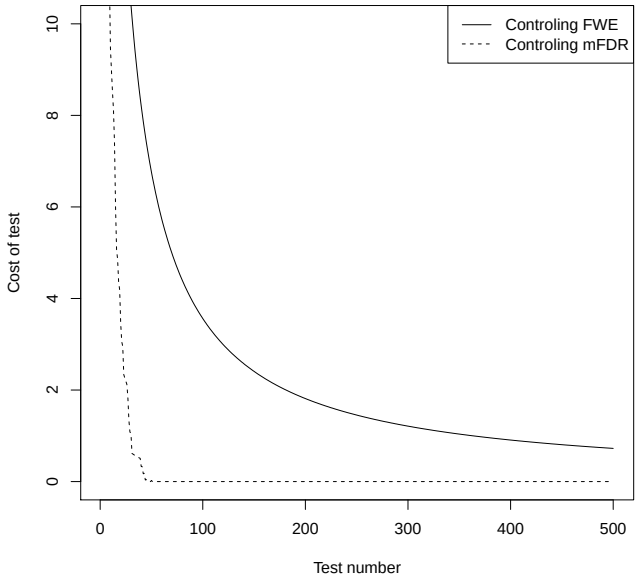

Figure 2: Average cost curves for running tests on normal means with power 0.95 at effect size 0.1 for a database starting at 2000 samples. The cost of each test is the number of additional samples (or cost of acquiring them) required as payments for running it. The full line represents costs when controlling FWER and the broken line are expected costs when controlling mFDR with 10% of tests having true effect (that is, each test is randomly assigned to be a false null with probability 0.1). The specific values depend on the parameters of the problem but the important and general observations are that costs decrease with time, and that controlling mFDR results in much lower costs, that can vanish if good hypotheses (i.e., false nulls) are occasionally tested. See Methods for description of calculations.