

# Is Replication the Gold Standard for Validating Genome-Wide Association Findings?

Yong-Jun Liu<sup>1</sup>, Christopher J. Papasian<sup>1</sup>, Jian-Feng Liu<sup>1</sup>, James Hamilton<sup>1</sup>, Hong-Wen Deng<sup>1,2,3\*</sup>

**1** School of Medicine, University of Missouri - Kansas City, Kansas City, Missouri, United States of America, **2** The Key Laboratory of Biomedical Information Engineering of Ministry of Education, and Institute of Molecular Genetics, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, People's Republic of China, **3** Laboratory of Molecular and Statistical Genetics, College of Life Sciences Hunan Normal University, Changsha, Hunan, People's Republic of China

## Abstract

With the advent of genome-wide association (GWA) studies, researchers are hoping that reliable genetic association of common human complex diseases/traits can be detected. Currently, there is an increasing enthusiasm about GWA and a number of GWA studies have been published. In the field a common practice is that replication should be used as the gold standard to validate an association finding. In this article, based on empirical and theoretical data, we emphasize that replication of GWA findings can be quite difficult, and should not always be expected, even when true variants are identified. The probability of replication becomes smaller with the increasing number of independent GWA studies if the power of individual replication studies is less than 100% (which is usually the case), and even a finding that is replicated may not necessarily be true. We argue that the field may have unreasonably high expectations on success of replication. We also wish to raise the question whether it is sufficient or necessary to treat replication as the ultimate and gold standard for defining true variants. We finally discuss the usefulness of integrating evidence from multiple levels/sources such as genetic epidemiological studies (at the DNA level), gene expression studies (at the RNA level), proteomics (at the protein level), and follow-up molecular and cellular studies for eventual validation and illumination of the functional relevance of the genes uncovered.

**Citation:** Liu Y-J, Papasian CJ, Liu J-F, Hamilton J, Deng H-W (2008) Is Replication the Gold Standard for Validating Genome-Wide Association Findings? PLoS ONE 3(12): e4037. doi:10.1371/journal.pone.0004037

**Editor:** Thorkild I. A. Sorensen, Copenhagen University Hospital, Denmark

**Received:** June 8, 2008; **Accepted:** November 21, 2008; **Published:** December 29, 2008

**Copyright:** © 2008 Liu et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** Investigators of this work were partially supported by grants from NIH (R01 AR050496, R21 AG027110, R01 AG026564, P50 AR055081 and R21 AA015973). The study also benefited from National Natural Science Foundation of China (30570875), Xi'an Jiaotong University, Huo YingDong Education Foundation, Hunan Province, and the Ministry of Education of China.

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

\* E-mail: dengh@umkc.edu

## Introduction

With the advent of genome-wide association (GWA) studies, it is anticipated that major susceptibility variants for common human diseases/traits can be detected [1]. With successful application of GWA studies first appearing in 2005, to date more than 190 GWA studies have been published, reporting more than 410 SNPs showing strong evidences of association with various human complex diseases/traits, such as obesity, diabetes, coronary heart diseases, asthma, cancers, mental illness, and osteoporosis (a catalog of published GWA studies is summarized at the National Human Genome Research Institute (NHGRI) website at [www.genome.gov/26525384](http://www.genome.gov/26525384)). Some GWA findings were confirmed in subsequent independent replication studies. With the belief that replication should be used as a gold standard in high quality publications, the NCI-NHGRI Working Group proposed suggestions in the June 2007 issue of *Nature* on what constitutes replication of a genotype-phenotype association, and how best to achieve this [2]. A commonly used review criterion for a GWA study is whether it is accompanied by internal (e.g., within consortium) or external (independent studies by other research groups in different populations) replication evidence.

In this article, based on empirical and theoretical data, we show that, like conventional linkage scan and candidate gene association

studies, GWA results are *also* going to be very difficult to replicate even for powerful, well-designed studies. We demonstrate that 1) the probability of replication becomes smaller as the number of independent GWA studies increases if the power of individual replication studies is less than 100% (which is usually the case), and 2) statistically replicated findings are not necessarily true, although replication lessens the likelihood of the initial finding being false. We question whether it is necessary or sufficient to regard replication as the gold standard for defining a true susceptibility variant. We lastly discuss the usefulness of integrating evidence from multiple level/sources such as genetic epidemiology, functional genomics, proteomics, and molecular and cellular functional studies.

## Results

### 1) GWA Results Are Inherently Difficult to Replicate

In GWA studies, a finding is considered to be "replicated" if and only if there is an initial finding which did appropriately control for multiple testing with strong control of the family-wise type I error rate (FWER) and in a second independent sample this finding is re-detected in a confirmatory hypothesis testing setting [2]. In the following we show that the probability of replicating GWA results is inherently low. We also discuss some confounding factors that may further exacerbate the situation.

**a) The probability of replicating GWA results is inherently low.** For simplicity, we assume an ideal situation in which a gene-phenotype association exists and is identified by the initial GWA. We then examine the probability of replicating the association under three slightly different scenarios: 1) the initial GWA findings are followed by several subsequent independent replication studies (either GWA or focused regional analyses); 2) two or more independent GWA studies are conducted simultaneously; and 3) a mixture of  $N$  GWA ( $N < 3-4$ ) and  $R$  replication studies ( $R > 5-6$ ) on a smaller number of markers. The power calculation in each scenario was performed using the “Genetic Power Calculator” which is public available (<http://pngu.mgh.harvard.edu/~purcell/gpc/>). All the power calculations in the “Genetic Power Calculator” are based upon formula derived in Sham et al. [3].

**Scenario one.** Due to polygenic inheritance of complex diseases/traits, the chance of replicating a susceptibility variant is *much* lower than that of initially detecting it. This is because in a GWA study designed to identify quantitative trait loci (QTLs), it is usually easy to detect *one* of the QTLs, even if the effect of the detected QTL is small and the statistical power of the study is low. As a numerical demonstration, suppose there are 20 QTLs underlie a complex trait, each explaining 1% phenotypic variation. Assuming an ideal situation that the QTL effect size of 0.01, the QTL allele frequency of 0.20, the marker allele frequency of 0.30, linkage disequilibrium (LD) between the QTL and the marker of 0.80 ( $D'$ ), and the significance level  $\alpha = 5.0 \times 10^{-7}$ , a GWA study with a large sample of 4,000 unrelated subjects (commonly used in current GWA studies) may only have  $\sim 10\%$  power to detect a *specific* QTL among these 20 QTLs, but this GWA study has a much higher power of  $\sim 88\%$  ( $1 - (1 - 10\%)^{20}$ ) to detect *at least one* of the QTLs.

However, the probability of replicating the *specific* QTL detected in the initial GWA will generally be low in subsequent replication studies. For example, under the *same* ideal situation assumed above, an independent replication study with a sample size of 4,000 unrelated subjects will only have  $\sim 49\%$  power to replicate the initial finding at a much less stringent significance level of  $\alpha = 10^{-4}$ . Therefore, the *specific* QTL identified in the initial GWA can be replicated unless a much larger sample with much higher power is used. In practice it should be noted that the actual effect of the identified variant in an initial study is usually overestimated, a phenomenon called winner’s curse [4–6]. Thus, the actual required sample size for replication is even larger than the one estimated based on the effect size reported in the initial discovery study.

**Scenario two.** We assume that two independent GWA studies are performed simultaneously. The likelihood that a *specific* susceptibility variant can be found by both studies (i.e., replicated by each other) is:

$$P(A_1, A_2 | +) = P(A_1 | +) \cdot P(A_2 | +) = \text{power}(A_1) \cdot \text{power}(A_2) \quad (1)$$

where  $P(A_1, A_2 | +)$ , the probability of detecting the susceptibility variant by both studies, is the product of the power of each study [denoted as  $\text{power}(A_1)$  and  $\text{power}(A_2)$ ]. When both studies have 10% power, the probability of detecting the same variant is as low as 1% (i.e.,  $10\% \times 10\%$ ). This probability is increased to 64% when both studies have much higher power of 80%. However, when one study has a high power of 80% and the other has low power of 10%, the probability of detecting the same variant is still as low as 8% ( $10\% \times 80\%$ ).

In the case of more than two independent GWA studies, consistent replication across studies becomes increasingly

challenging and the probability of replication becomes diminishingly small unless each study has power close to 100%. Assuming that  $N$  independent GWA studies are performed, the probability of replicating a *specific* susceptibility variant by all the studies is

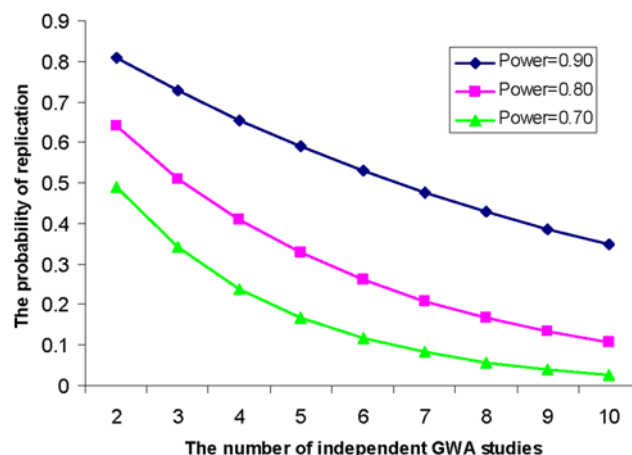
$$P(A_1, \dots, A_n | +) = \prod_{i=1}^n \text{power}_{A_i} \quad (2)$$

It can be seen that the likelihood of consistent replication by  $N$  individual studies depend on the statistical power of each study, and particularly, on the study of the *lowest* power. As the number of individual, independent studies increases, there is a dramatic decrease in the probability that the *same* variant will be consistently significant in all of those studies (shown in **Figure 1**).

A possible illustration for this scenario is four large, recently published GWA studies for type 2 diabetes [7–10]. The first published one was performed in a French case-control cohort [10]. The other three were performed, respectively, by the Finland-United States Investigation on NIDDM Genetics (FUSION) team [9], the Diabetes Genetics Initiative [8], and the Wellcome Trust Case-Control Consortium [7]. Each GWA identified a number of loci that may confer type 2 diabetes risk. However, only a limited few loci (e.g., *TCF7L2* and *SLC30A8*) showed significant associations across ALL the four studies. Notably, neither of these loci achieved a genome-wide significance level (a  $P$  value of  $10^{-7}$  [11]) in all the four studies.

**Scenario three.** This scenario represents a hybrid of scenarios one and two. We assume that there are  $N$  independent GWA and  $R$  replication studies. Without loss of generality, let’s consider a situation, where two independent GWA ( $N = 2$ ) are followed by six replication studies ( $R = 6$ ) on a smaller number of markers.

Suppose there are 20 QTLs, each explaining 1% phenotypic variance. As shown in Scenario one, under an ideal situation, a GWA study with a sample size of 4,000 may have only  $\sim 10\%$  power to detect a *specific* QTL among these 20 QTLs, but the study may have a high power of  $\sim 88\%$  ( $1 - (1 - 10\%)^{20}$ ) to detect *at least one* of these 20 QTLs. The probability that the same QTL can be identified by both initial GWA studies is only  $\sim 8.8\%$  ( $10\% \times 88\%$ ). To achieve 80% power to replicate this association in follow-up studies (assuming  $D' = 0.80$ , QTL allele frequency of 0.20, marker allele frequency of 0.30,  $\alpha$  level of  $10^{-4}$ , no



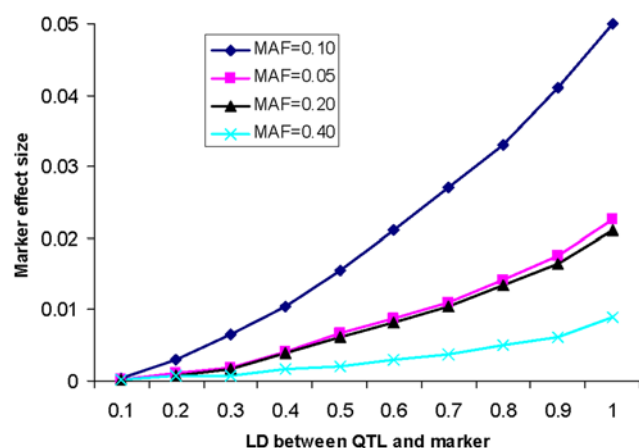
**Figure 1. The probability of replication in different numbers of independent GWA studies.** We assume all the studies have the same power (0.90, 0.80, and 0.70) to detect a specific genotype-phenotype association.

doi:10.1371/journal.pone.0004037.g001

between-study heterogeneity or other potential biases), a larger sample of at least 6,000 subjects is needed. Even if replication is achieved, there is still a possibility that the finding is false positive (shown below in section “Are Replicated GWA Findings Always True?”).

**b) Influence of LD and allele frequency difference.** GWA is an “indirect” approach testing association between markers and diseases/traits of interest. Thus, not only the frequency of susceptibility variants but also the markers may affect the likelihood of detecting/replicating associations. The influence of LD and allele frequency difference on association studies has been demonstrated for diseases [12–14], but their quantitative effects on *quantitative* traits (e.g., BMI) have not been studied. To further illustrate the principles for quantitative traits, we show how these population-specific characteristics (i.e., marker and susceptibility allele frequencies, and LD strength) may affect study power. We consider this in terms of marker effect size, which we can estimate in practical studies.

Let’s consider a diallelic QTL ( $Q/q$ ) and a diallelic marker ( $M/m$ ), with frequencies of  $P_Q$ ,  $P_q$ ,  $P_M$ , and  $P_m$ , respectively. Assuming LD ( $D'$ ) between the QTL and the marker ranges from 0.1 to 1.0, we can calculate the marker effect size (see **Appendix S1** for detail). **Figure 2** shows how the measured marker effect sizes are influenced by frequencies of the QTL and marker and LD between them. It can be seen that the power of a GWA is greatest when allele frequencies of the QTL and the marker match. Discrepancies in allele frequency between the QTL and the marker may reduce power, often dramatically as the magnitude of this power reduction increases as the discrepancy in allele frequency increases. When susceptibility variant and marker allele frequencies match, the marker effect size is approximately proportional to the QTL effect size multiplied by the  $D'$  value between the two loci. Since allele frequencies of the QTLs and markers and LD strength may vary across populations [15–17], the power of GWA studies in different study populations may have different power to detect a QTL, even if the sample sizes are the same. A situation that may make replication even more difficult is long-range LD or cross-chromosome LD which may exist in the human genome [18]. Intuitively, such long range LD may have greater variation between different populations.



**Figure 2. How marker effect size (y-axis) is determined by QTL effect size, marker allele frequency (MAF) and QTL allele frequency, and extent of LD between marker and QTL.** We assume the QTL is under additive inheritance with MAF of 0.10 and effect size of 0.05 and the LD is measured by  $D'$  (x-axis). doi:10.1371/journal.pone.0004037.g002

**c) Confounding factors.** Diversity in subject ascertainment and study design across studies may significantly impact GWA replication [11]. In addition, confounding factors, such as population stratification, genetic heterogeneity, environmental factors, and interactions between genetic and environmental factors (which may vary with different populations and environments), may reduce the chance of GWA replication. Most of these factors have been well recognized and evaluated in association studies for candidate genes [19]. Here we only briefly review these factors with regard to GWA studies.

Population stratification/admixture may mask, change or even result in apparently reversed genetic effect of genes underlying complex disease/traits. The available methods addressing/controlling this problem may have potential limitations [20–23]. For example, the Genome Control method [20] assumes that the degree of population differentiation is the same throughout the human genome, which may not be the case. The performance of the Structured Association method [23] is highly dependent upon the amount and informativeness of ancestry informative markers (AIMs) [24], and the accuracy of inferred individual ancestry is sensitive to the number of pre-assigned subpopulations, which may not always be satisfactorily resolved [25]. In addition, the use of the Structured Association method for GWA studies is limited due to its intensive computational cost for large data sets. Principal component analysis [22] is a recently developed method specifically for GWA; however, it may lead to incomplete stratification correction when the number of markers is less than 20,000 [26].

Genetic heterogeneity is a common phenomenon in human diseases. A disease could be caused by different susceptibility variants in different populations. For a specific susceptibility variant, there may also exist genuine diversity of its genetic effect in different populations [27]. The allele frequency of a susceptibility variant could be different across different populations. Since the population genetic effect of a susceptibility variant depends on its allele frequency (**Figure 2**), it should not be surprising that a significant association identified in one population cannot be found in another. A recent study estimated the required sample size to replicate an association finding with different amounts of between-study heterogeneity [28]. The authors concluded that: 1) if between-study heterogeneity reaches certain thresholds, it may not be practically possible to consistently replicate some true associations, no matter how large the studies are, and 2) replication sample sizes of 40,000 subjects or even larger are essential for generating sufficient power to replicate an association of small or modest effect size [28].

Studies have shown that LD patterns vary substantially among different populations [29–32]. This differential LD (or LD heterogeneity) may cause nonreplication of GWA findings across studies in different populations [27]. LD heterogeneity may exist even in populations of the same ethnic group such as Caucasians of European ancestry [30,33,34], although data accumulated from recent GWA studies [35–37] tend to show that, between Caucasian populations, the LD is not very heterogeneous. The diverse intermarker LD may affect the probability of replicating a gene or region and this should be considered when designing a replication study, especially with regard to marker-selection strategy. Clarke et al. [38] showed that when a region of high intermarker LD is tested to replicate an initial finding that is only weak association with a disease, the “local” approach that involves both the originally significant markers and others in the same regions is a good strategy. Otherwise, the most powerful and efficient strategy for replication involves testing only the initially identified variants [38].

Populations of different or even similar cultures may have different exposures to environmental factors. For example, lifestyles such as diet, smoking, alcohol drinking, nutritional status, and exercises may have significant influence on human body fat (a focal phenotype for obesity research). These factors are sometimes difficult to be assessed and quantified accurately and their actual influence on body fat variation might not be judiciously accounted for in statistical analyses, although there is a debate on whether the statistical adjustment for environmental factors is necessary [39].

## 2) Are Replicated GWA Findings Always True?

In the field a common practice is that a significant GWA finding is considered to be “true” if replicated by several subsequent studies (not necessarily all). However, are replicated GWA findings always true? Some Bayesian based approaches (such as positive predictive value (PPV) to be discussed below) or complementary approaches (such as Bayesian False-Positive Report Probability) [40,41] may help assess the credibility of association findings. In the following, we assess the probability of the replicated association being “true” using a Bayesian approach - PPV [42].

For an association finding, the PPV can be estimated by

$$\begin{aligned} PPV = P(+|A^+) &= \frac{P(A^+|+)P(+)}{P(A^+|+)P(+)+P(A^+|-)P(-)} \\ &= \frac{(1-\beta)P(+)}{(1-\beta)P(+)+p[1-P(+)]} \end{aligned} \quad (3)$$

where ‘+’ and ‘-’ denote the presence and absence of association, respectively.  $A^+$  denotes rejection of the null hypothesis (i.e., a significant association is detected).  $P(A^+|+)=1-\beta$  is the statistical power and  $P(A^+|-)=p$  is the type I error rate.  $P(+)$  is the prior (pre-study) probability that the association is true. Similarly, we can use negative predictive value (NPV) to evaluate the probability of null hypothesis being true (i.e., no association) when a follow-up study does not find the association. NPV can be estimated by

$$\begin{aligned} NPV = P(-|A^-) &= \frac{P(A^-|-)P(-)}{P(A^-|-)P(-)+P(A^-|+)P(+)} \\ &= \frac{(1-p)[1-P(+)]}{(1-p)[1-P(+)]+\beta\cdot P(+)} \end{aligned} \quad (4)$$

where  $A^-$  denotes acceptance of the null hypothesis (i.e., no association is detected).

Supposing a significant association is identified in the initial GWA, the probability of the identified association being true is

$$P_1(+) = PPV_1 = P(+|A_1^+) = \frac{(1-\beta)P_0(+)}{(1-\beta)P_0(+)+p[1-P_0(+)]} \quad (5)$$

where  $P_1(+)$  is the PPV of the initial GWA.

Further suppose that  $n$  independent subsequent GWA studies are performed. If the  $k$ th study ( $k \leq n$ ) replicates the initial finding (i.e., the null hypothesis is rejected), the probability of the identified association being true is

$$\begin{aligned} P_k(+) &= PPV_k = P(+|A_k^+) \\ &= \frac{(1-\beta)P_{k-1}(+)}{(1-\beta)P_{k-1}(+)+p[1-P_{k-1}(+)]} \end{aligned} \quad (6)$$

Alternatively, if the  $k$ th study does not replicate the initial finding (i.e., the null hypothesis is accepted), the probability the association being true is

$$\begin{aligned} P_k(+) &= 1 - NPV_k = 1 - P(-|A_k^-) \\ &= 1 - \frac{(1-p)[1-P_{k-1}(+)]}{(1-p)[1-P_{k-1}(+)]+\beta\cdot P_{k-1}(+)} \\ &= \frac{\beta\cdot P_{k-1}(+)}{(1-p)[1-P_{k-1}(+)]+\beta\cdot P_{k-1}(+)} \end{aligned} \quad (7)$$

From Equations 5–7, it can be seen that the probability of a replicated GWA finding being true is determined by pre-study prior probability, statistical power, and  $p$  value of the test. Note, here the prior probability changes by incorporating the PPV of previous studies.

The above Equations are given without considering potential biases. Here biases refer to those factors (due to study design, data quality, statistical analysis, or result presentation) that tend to produce misleading research results. Let’s quantify bias by  $u$  - the proportion of probed analyses that would not have been GWA findings in an unbiased study, but are nevertheless reported as such, because of bias [42]. By taking into account bias  $u$ , the probability of an identified association being true can be calculated as

$$\begin{aligned} P_1(+) &= PPV_1 = P(+|A_1^+) \\ &= \frac{(1-\beta+u\beta)P_0(+)}{(1-\beta+u\beta)P_0(+)+[p(1-u)+u][1-P_0(+)]} \end{aligned} \quad (8)$$

If  $n$  independent subsequent GWA studies are performed and the  $k$ th study ( $k \leq n$ ) replicates the initial finding, the probability of the association being true is

$$\begin{aligned} P_k(+) &= PPV_k = P(+|A_k^+) \\ &= \frac{(1-\beta+u\beta)P_{k-1}(+)}{(1-\beta+u\beta)P_{k-1}(+)+[p(1-u)+u][1-P_{k-1}(+)]} \end{aligned} \quad (9)$$

As a numerical illustration, we assume that a number of sequential GWA studies are performed using 500 K SNP arrays, among which 20 SNPs are susceptibility variants. Then the prior (pre-study) probability of finding at least one of the susceptibility variants is  $P_0(+)=20/500,000=4.0 \times 10^{-5}$ , which is quite low compared to hypothesis-driven studies with prior knowledge or evidence of the tested association. We next consider an actual situation where both positive and negative results are involved. We assume that an initial GWA study identified a significant genotype-phenotype association, which is followed up by one to six subsequent independent studies. We show in **Table 1** the PPV values corresponding to various situations where follow-up studies replicated the initial finding. From Table 1, we can see that if three among four follow-up studies replicate the initial association and the fourth does not, the PPV of the identified association is only  $\sim 0.31$ . This means the chance of being true is only 31% and the chance that this replicated finding is false is still as high as 69%. If assuming a bias  $u=0.1$ , the PPV decreases dramatically to as low as 0.006. As another example, if four among six follow-up studies replicate an initial association and two other follow-up studies do not, the PPV of the identified association is only 0.46. While PPV



**Table 1.** The *PPVs* corresponding to various situations where different number of follow-up studies replicated the initial finding.

GWA studies		PPV						
Initial GWA		0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.00072
Replication studies	1	0.012	<b>7.58E-05</b>	<b>7.58E-05</b>	<b>7.58E-05</b>	<b>7.58E-05</b>	<b>7.58E-05</b>	<b>7.58E-05</b>
	2	0.189	0.001	<b>7.98E-06</b>	<b>7.98E-06</b>	<b>7.98E-06</b>	<b>7.98E-06</b>	<b>7.98E-06</b>
	3	0.807	0.023	0.0001	<b>8.40E-07</b>	<b>8.40E-07</b>	<b>8.40E-07</b>	<b>8.40E-07</b>
	4	0.986	0.306	0.002	1.51E-05	<b>8.84E-08</b>	<b>8.84E-08</b>	<b>8.84E-08</b>
	5	0.999	0.888	0.044	0.0002	1.59E-06	<b>9.31E-09</b>	<b>9.31E-09</b>
	6	0.999	0.993	0.455	0.004	2.86E-05	1.67E-07	<b>9.8E-10</b>

PPV: positive predictive value.

Bold represents the follow-up study that does not replicate the initial finding.

doi:10.1371/journal.pone.0004037.t001

increases with the increasing number of positive replication studies, each non-replicated study may dramatically decrease the *PPV* value. The above *PPV* calculation was based on an assumption that all the studies have high statistical power of 90%, which is often higher than that achieved in actual studies.

It should be noted that the experiment-wise  $\alpha$  level of 0.05 is used in the above *PPV* analysis. In GWA studies, to account for multiple comparison by testing hundreds of thousands SNPs, a point-wise *P* value of  $10^{-7}$  for a genotype-phenotype association is considered to be statistically significant [11]. However, in *PPV* analysis, only the experiment-wise *P* values that are adjusted for multiple testing should be used. Otherwise, even point-wise *P* values on the order of  $10^{-2}$ – $10^{-5}$  achieved in an initial GWA can yield high *PPV* if directly used in *PPV* analyses without correction for multiple testing.

## Discussion

### Integrating Evidence of Multiple Levels/Sources

By discussing the difficulties of replication, we do not intend to depreciate the GWA approach and subsequent replication efforts. This is an evolving area and our understanding of GWA is improving from the knowledge gained and challenges that remain. The genetic community has had some useful discussions on design, implementation, best practice and interpretation of GWA studies [11]. Even with the difficulty of replication notwithstanding, variants with relatively large effects may be identified and replicated in powerful GWA studies [11]. A potential practical example is the *FTO* gene, whose association with obesity and related phenotypes was identified in three independent GWA studies in French [35], German [36], and British populations [37]. This *FTO* gene is responsible for 1% of the total heritability of obesity [43]. Meta-analysis is a useful tool for synthesizing data and exploring potential heterogeneity [44,45]. Joint (meta) analysis of data from comparable GWA studies may increase the power of gene identification when individual GWA studies are underpowered. However, meta-analysis may not always be ideal as it may suffer from potential problems such as between-study heterogeneity and bias (e.g., selective publication) [46].

From systems biology perspective, GWAs are studies at the DNA level. At best, GWA studies could, in conjunction with fine mapping efforts, identify the implicated variant(s) at the level of molecular markers (e.g., SNPs). Although GWA studies may implicate a gene(s) as a factor contributing to a disease/trait, it cannot tell anything about *how* the gene(s) contribute to the disease/trait. Similarly, bioinformatics tools, while useful, may only infer some of the functions of the identified markers/genes.

Functional genomic studies, including gene expression studies at the RNA level and proteomics studies at the protein level, may provide useful complementary information to GWA studies. For example, functional genomic studies may unravel critical information about the regulation of gene activity under various conditions that may contribute to our knowledge of molecular and genetic mechanisms influencing disease development. GWAs, gene expression and proteomics studies, individually, have shown some successes in identifying genes for complex diseases. However, each may be prone to false positive/negative findings, partially due to multiple testing in genomic approaches, statistical power (partially associated with conservative procedures accounting for multiple testing), and the biological complexity of gene expression and genetic etiology. Gene expression is regulated simultaneously and interactively at all the three levels, i.e., DNA, RNA, and protein levels. Gene expression may be regulated at the level of DNA, RNA, or protein and there is often interaction between regulatory controls at these different levels. Hence, a genomic convergence or systems biology based approach that integrates the information from GWA studies, gene expression and proteomics may facilitate the identification of key pathways that are globally involved in the pathogenesis of the disease and/or interactive factors acting at different levels of disease-gene expression [47]. It should be noted that genomic convergence may have its own limitations. Genes identified at the DNA level in GWA studies may escape confirmation/replication in studies of RNA (microarray) or protein (proteomics), partially because of inherent differences in these experimental approaches and complex regulation of gene functions involved, even when those genes actually contribute to disease development.

Gene expression levels can be used as quantitative traits in traditional linkage or association studies, and the identified loci are termed expression QTLs (eQTLs) [48]. Given that most common human diseases are outcome of a complex interaction between many genetic loci and the environment, there are obvious advantages to studying the genetics of gene expression in cells that represent the *in vivo* state. For example, a recent study analyzed the expression of 23,720 transcripts in large population-based blood and adipose tissue cohorts for various obesity related traits [49]. A core network module in humans and mice was identified that is enriched for genes involved in the inflammatory and immune response and was found to be causally associated to obesity-related traits [49].

Ultimately, the functional relevance of the identified variant(s) should be confirmed by *in vivo* or *in vitro* studies. These functional studies, however, also face challenges, as results of *in vivo* and *in vitro*

studies might not be consistent, and findings obtained in animal models may not necessarily translate into those in humans. How to best integrate information of multiple levels/sources in gene identification and functional studies remain a topic for open discussion. This is an area that requires substantial efforts from biologists and clinicians, statistical geneticists, bioinformaticians, and epidemiologists.

In summary, identification of genes underlying human complex diseases is challenging. Replication, while important and valuable, is difficult to achieve and may not be sufficient or necessary for validating GWA findings. Additional information from other lines of evidence, such as detailed molecular mechanistic studies and genomic convergence, may be useful for validating and illuminating the functional relevance of genes identified in GWA studies.

## Methods

The power calculation was performed using the “Genetic Power Calculator” (<http://pngu.mgh.harvard.edu/~purcell/gpc/>). All the power calculations in the “Genetic Power Calculator” are based upon formula derived in Sham et al. [3].

## References

- Couzin J, Kaiser J (2007) Genome-wide association. Closing the net on common disease genes. *Science* 316: 820–822.
- Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, et al. (2007) Replicating genotype-phenotype associations. *Nature* 447: 655–660.
- Sham PC, Cherny SS, Purcell S, Hewitt JK (2000) Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Genet* 66: 1616–1630.
- Goring HH, Terwilliger JD, Blangero J (2001) Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet* 69: 1357–1369.
- Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. *Genet Med* 4: 45–61.
- Zollner S, Pritchard JK (2007) Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am J Hum Genet* 80: 605–615.
- Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661–678.
- Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316: 1331–1336.
- Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, et al. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316: 1341–1345.
- Sladek R, Rocheleau G, Rung J, Dina C, Shen L, et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 445: 881–885.
- McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9: 356–369.
- Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* 56: 18–31.
- Moskvina V, O'Donovan MC (2007) Detailed analysis of the relative power of direct and indirect association studies and the implications for their interpretation. *Hum Hered* 64: 63–73.
- Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. *Nat Rev Genet* 5: 89–100.
- Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, et al. (2002) The structure of haplotype blocks in the human genome. *Science* 296: 2225–2229.
- Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, et al. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* 307: 1072–1079.
- Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, et al. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294: 1719–1723.
- Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, et al. (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* 68: 191–197.
- Colhoun HM, McKeigue PM, Davey SG (2003) Problems of reporting genetic associations with complex outcomes. *Lancet* 361: 865–872.
- Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55: 997–1004.
- Kohler K, Bickeboller H (2006) Case-control association tests correcting for population stratification. *Ann Hum Genet* 70: 98–115.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38: 904–909.
- Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 65: 220–228.
- Tsai HJ, Choudhry S, Naqvi M, Rodriguez-Cintron W, Burchard EG, et al. (2005) Comparison of three methods to estimate genetic ancestry and control for stratification in genetic association studies among admixed populations. *Hum Genet* 118: 424–433.
- Setakis E, Stirnadel H, Balding DJ (2006) Logistic regression protects against population structure in genetic association studies. *Genome Res* 16: 290–296.
- Epstein MP, Allen AS, Satten GA (2007) A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet* 80: 921–930.
- Terwilliger JD, Hiekkalinna T (2006) An utter refutation of the “Fundamental Theorem of the HapMap”. *Eur J Hum Genet* 14: 426–437.
- Moonesinghe R, Khoury MJ, Liu T, Ioannidis JP (2008) Required sample size and nonreplicability thresholds for heterogeneous genetic associations. *Proc Natl Acad Sci U S A* 105: 617–622.
- Evans DM, Cardon LR (2005) A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *Am J Hum Genet* 76: 681–687.
- Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, et al. (2005) Linkage disequilibrium patterns vary substantially among populations. *Eur J Hum Genet* 13: 677–686.
- Liu N, Sawyer SL, Mukherjee N, Pakstis AJ, Kidd JR, et al. (2004) Haplotype block structures show significant variation among populations. *Genet Epidemiol* 27: 385–400.
- De LV, Isaac H, Collins A, Scafe CR, Halldorsson BV, et al. (2005) The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. *Genome Res* 15: 454–462.
- Mueller JC, Lohmussaar E, Magi R, Remm M, Bettecken T, et al. (2005) Linkage disequilibrium patterns and tagSNP transferability among European populations. *Am J Hum Genet* 76: 387–398.
- Lundmark PE, Liljedahl U, Boomsma DI, Mannila H, Martin NG, et al. (2008) Evaluation of HapMap data in six populations of European descent. *Eur J Hum Genet* 16: 1142–1150.
- Dina C, Meyre D, Gallina S, Durand E, Korner A, et al. (2007) Variation in FTO contributes to childhood obesity and severe adult obesity. *Nat Genet* 39: 724–726.
- Scuteri A, Sanna S, Chen WM, Uda M, Albai G, et al. (2007) Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet* 3: e115.
- Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316: 889–894.

## Supporting Information

**Appendix S1** Determination of marker effect size by QTL effect size, allele frequency differences, and the LD between marker and QTL

Found at: doi:10.1371/journal.pone.0004037.s001 (0.08 MB DOC)

## Acknowledgments

The authors are grateful to Ms. Jillian Richwine for her help on manuscript editing.

## Author Contributions

Conceived and designed the experiments: YL HWD. Analyzed the data: YL JL. Contributed reagents/materials/analysis tools: YL. Wrote the paper: YL CP JL JPH HWD.

38. Clarke GM, Carter KW, Palmer LJ, Morris AP, Cardon LR (2007) Fine mapping versus replication in whole-genome association studies. *Am J Hum Genet* 81: 995–1005.
39. Smith GD, Lawlor DA, Harbord R, Timpson N, Day I, et al. (2007) Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Med* 4: e352.
40. Wacholder S, Chanock S, Garcia-Closas M, El GL, Rothman N (2004) Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* 96: 434–442.
41. Wakefield J (2007) A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet* 81: 208–227.
42. Ioannidis JP (2005) Why most published research findings are false. *PLoS Med* 2: e124.
43. Dina C (2008) New insights into the genetics of body weight. *Curr Opin Clin Nutr Metab Care* 11: 378–384.
44. Evangelou E, Maraganore DM, Ioannidis JP (2007) Meta-analysis in genome-wide association datasets: strategies and application in Parkinson disease. *PLoS ONE* 2: e196.
45. Ioannidis JP, Patsopoulos NA, Evangelou E (2007) Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS ONE* 2: e841.
46. Munafò MR, Flint J (2004) Meta-analysis of genetic association studies. *Trends Genet* 20: 439–444.
47. Lusis AJ, Attie AD, Reue K (2008) Metabolic syndrome: from epidemiology to systems biology. *Nat Rev Genet* 9: 819–830.
48. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* 37: 710–717.
49. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, et al. (2008) Genetics of gene expression and its effect on disease. *Nature* 452: 423–428.