Fine Mapping versus Replication in Whole-Genome Association Studies

Geraldine M. Clarke, Kim W. Carter, Lyle J. Palmer, Andrew P. Morris, and Lon R. Cardon

Association replication studies have a poor track record and, even when successful, often claim association with different markers, alleles, and phenotypes than those reported in the primary study. It is unknown whether these outcomes reflect genuine associations or false-positive results. A greater understanding of these observations is essential for genomewide association (GWA) studies, since they have the potential to identify multiple new associations that that will require external validation. Theoretically, a repeat association with precisely the same variant in an independent sample is the gold standard for replication, but testing additional variants is commonplace in replication studies. Finding different associated SNPs within the same gene or region as that originally identified is often reported as confirmatory evidence. Here, we compare the probability of replicating a gene or region under two commonly used marker-selection strategies: an "exact" approach that involves only the originally significant markers and a "local" approach that involves both the originally significant markers and others in the same region. When a region of high intermarker linkage disequilibrium is tested to replicate an initial finding that is only weak association with disease, the local approach is a good strategy. Otherwise, the most powerful and efficient strategy for replication involves testing only the initially identified variants. Association with a marker other than that originally identified can occur frequently, even in the presence of real effects in a low-powered replication study, and instances of such association increase as the number of included variants increases. Our results provide a basis for the design and interpretation of GWA replication studies and point to the importance of a clear distinction between fine mapping and replication after GWA.

Genomewide association (GWA) studies are now under way, involving hundreds of thousands of genetic markers genotyped in thousands of individuals.[1–3] The hope for such studies is that they will identify major loci, although the expectation is that most findings will comprise smaller-effect variants that appear to be more likely to occur than by chance alone, but which initially are not exceptional enough to be unambiguously related to the outcomes.[4–6] Meta-analyses and results of initial genomewide studies support this pattern of effects.[7–19] Thus, there soon will be thousands of possible disease loci for which some indicative evidence emerges from GWA studies, but which will then require further scrutiny for validation.

Independent replication, the process by which validation of study results may be achieved, has long been the strategy of choice to validate initial reports of genetic association. Unfortunately, such studies are regarded as one of the weakest aspects of human genetics, achieving a disproportionately small number of successful outcomes despite tens of thousands of attempts.[20–22] Confusingly, even when success is declared for replication, many of the studies show seemingly implausible patterns in which different markers, alleles, and phenotypes are found to be associated in the initial and subsequent studies.[23] In some cases, the disease-predisposing risk allele in the primary study is reported as the "protective" allele in the replication study.[23,24] The contradictory results are then ascribed to undetected genetic, allelic, phenotypic, and population heterogeneity.[25] A brief history and description of prospects for replication studies of complex traits has been presented recently by Chanock et al.[26]

In theory, an additional significant association with exactly the same allele in an independent sample is the gold standard for replication. However, this is not often the practice of association researchers. Instead, so-called replication studies often comprise a hybrid design involving rigorous testing of the same markers (exact replication) plus fine mapping of additional loci in the same region. There are several reasons why such a design might be employed. First, the indirect nature of whole-genome scans means that other markers in an association region may be more strongly correlated with the disease variant(s) and thereby offer greater statistical power to detect an association. This possibility has motivated the development of "gene-based" association tests that expressly allow for marker and allelic variability in the patterns of allelic association.[27] Additional variants are also pursued in replication studies, because of their location in coding or promoter regions, low levels of linkage disequilibrium (LD), marker gaps in the initial coverage, allele-frequency hypotheses, external biological information, or prior assumptions. This desire to include additional markers,

From the Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom (G.M.C.; A.P.M.; L.R.C.); Centre for Genetic Epidemiology, Western Australia Institute for Medical Research, Nedlands (K.W.C.; L.J.P.); and UWA Centre for Medical Research, University of Western Australia, Perth (K.W.C.; L.J.P.)

Received February 14, 2007; accepted for publication July 25, 2007; electronically published September 19, 2007.

Address for correspondence and reprints: Dr. Lon R. Cardon, Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, United Kingdom. E-mail: lon.cardon@well.ox.ac.uk

Am. J. Hum. Genet. 2007;81:995–1005. © 2007 by The American Society of Human Genetics. All rights reserved. 0002-9297/2007/8105-0012$15.00
DOI: 10.1086/521952

combined with the difficulties in irreproducibility of exact replication, has led many investigators to use what we refer to as "local" replication, whereby a subsequent significant association observed within the same gene or region in an independent sample is taken as confirmation of the role of that gene or region in the etiology of the disease of interest.[23] The differences between exact and local replication strategies are depicted in figure 1.

Note that this form of replication differs from "replication" in multistage GWA studies or in situations where data from multiple sample resources can be combined for greater power of initial detection (or, more commonly, where data from a single resource is not split into separately analyzed pieces to "save" some samples for replication).[28] Such strategies focus on initial detection and are thus hypothesis generating. Replication in the present context is of the traditional hypothesis-testing form, used broadly by different investigators with different samples and study designs. These types of replication studies aim to independently verify initial association reports and to obtain accurate effect-size estimates, regardless of the designs used in the primary study to detect the effects. The eventual utility, robustness, and acceptance of the GWA approach will depend in part on the correct design, execution, and interpretation of such studies. Here, we formally compare the exact and local replication strategies in the presence of a real association, to assess the performance of each approach under different conditions and to determine their practical utility as follow-up designs for GWA.

Material and Methods

Study Designs

We refer to the putatively associated markers in the primary study as the "representative" markers. The exact replication strategy involves testing only the representative marker(s) from the original study; the local replication strategy involves testing the representative marker(s) plus other markers that may or may not have been genotyped in the primary study but that are selected because of LD patterns, candidate genes, gaps in initial coverage, or other hypotheses of the investigator (fig. 1). For both designs, we assume that representative markers are used to identify K regions of interest, each of which contains one or more genuine disease-susceptibility loci. Note that the actual disease loci are not necessarily genotyped in either the primary or the replication studies, but they are correlated with markers in the regions. The replication study then involves genotyping a representative marker and $M-1$ biallelic markers in each region. Markers are assumed to be in LD, measured by r^2, with the causal alleles. If $r^2>0$, then a marker is defined to be true. We assume that m ($m \leq M$) of the markers tested in each region are true. We declare replication to be successful in a region when one or more of the tested markers exceed the significance criteria of both the initial and the replication studies. In all cases, we assume that the replication sample is independent of the initial sample but is drawn from the same population and is matched with respect to phenotype, covariate, and ascertainment effects. Note that, under these assumptions, the conditional probability of replication is simply the probability of a successful association in the second test; here, we report the joint probability of replication given the presence of a real association.

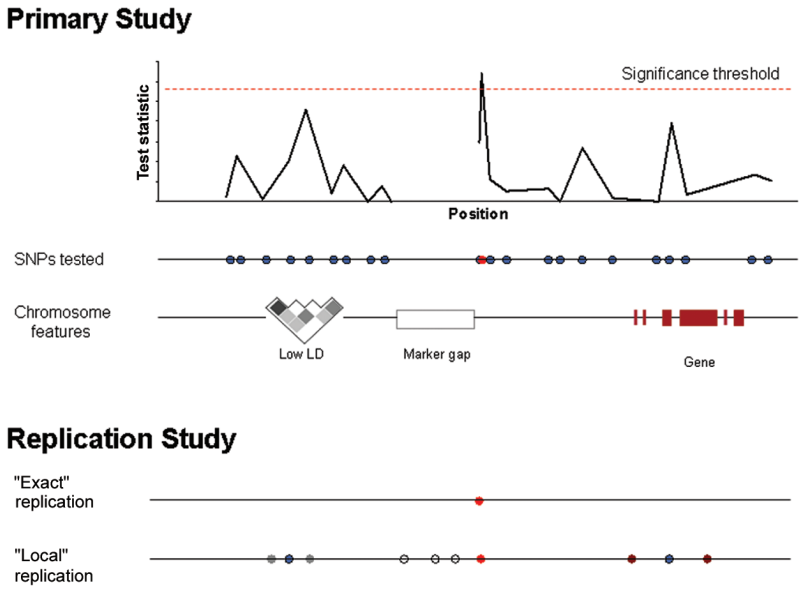

Figure 1. Exact and local replication strategies. Exact strategies involve testing only those markers that exceed some significance threshold in the primary study, whereas local replication studies involve testing additional markers on the basis of genomic information, such as LD patterns, marker gaps, and gene locations (*shown*) or other prior hypotheses. In the local strategy, the markers tested may include some loci that were not deemed "significant" in the initial study, as well as new SNPs that were not tested initially at all.

Analytical Probabilities for Local versus Exact Designs

We consider single-marker tests of association and define a marker to be successful in a given study if it exceeds the significance criteria of that study. Let $P_1(i)$, $i = 1, \ldots, L$, be the event that marker i is successful in the primary study. Suppose that a region containing a single representative marker and $M - 1$ additional markers is identified, and markers are genotyped as part of a replication study. Let $P_2(j)$, $j = 1, \ldots, M$, be the event that marker j is successful in the replication study. Let $R(i)$ be the event that a successful replication occurred at locus i—that is, that representative marker i was successful in the initial study, and at least one other marker in the region was successful in the replication study. Then, a general expression for $R(i)$ is

$$\Pr[R(i)] = \Pr[I_{P_1(i)} = 1]\{1 - \Pr[S_2(M) = 0]\} ,$$

where $\Pr[S_2(M) = 0] = \Pr[I_{P_2(1)} = 0, I_{P_2(2)} = 0, \ldots, I_{P_2(M)} = 0]$ is the joint probability of no successful replication at any of the markers in the replication study and is an indicator function that takes the value 1 if it is true and 0 otherwise.

Given case and control allele frequencies at the causal locus, sample sizes, and the ratio of cases to controls, the marginal probability of a single marker being successful in a standard χ^2 test of association is dependent only on the intermarker LD between the alleles at the marker and the causal locus.[29] To accommodate the testing of multiple markers, we need to consider the intermarker LD between alleles at each pair of test markers, since the LD relationships at one pair of markers generate constraints on the range of LD at other correlated markers.[30] A marker is defined to be true if it has a real association with a causal locus; otherwise, it is false. Here, we assume zero intermarker LD between any pairs of markers in the region that include at least one false marker, an assumption that is not expected to affect the overall replication probabilities but that simplifies calculation of probabilities and multiple-testing corrections. Suppose that m of the M markers in the region are true. For simplicity of notation, we arbitrarily order the true markers as $1, \ldots, m$ and the false markers as $m + 1, \ldots, M$. Then,

$$\Pr[R(i)] = \Pr[I_{P_1(i)} = 1]\left\{1 - \Pr[S_2(m) = 0] \prod_{j=m+1}^{M} \Pr[I_{P_2(j)} = 0]\right\} . \quad (1)$$

Instead of considering specific LD patterns, which vary widely and sometimes unpredictably between chromosome regions and samples,[27] we examine the upper and lower bounds that delineate the range of possible replication outcomes for different LD patterns.

In regions of perfect intermarker LD between pairs of true markers ($r^2_{j,k} = 1$; $j,k \leq m$; $j \neq k$), the true markers collapse to a single locus, providing a lower bound for $\Pr[R(i)]$:

$$B_l \equiv \Pr[R(i)] \geqslant \Pr[I_{P_1(i)} = 1]\left\{1 - \Pr[I_{P_2(1)} = 0] \prod_{j=m+1}^{M} \Pr[I_{P_2(j)} = 0]\right\} , \quad (2)$$

where we have arbitrarily used the probability of success at marker 1 because, when $r^2_{jk} = 1$ for all m markers, all the true markers in the replication study must have the same probability of success. Note that, in the case of the exact replication strategy, $m = M$, so expression (2) reduces to the probability of initially detecting and replicating a trait association with marker 1. In cases of local replication, we make a Bonferroni correction with $\alpha' = \alpha/K(M - m + 1)$. Note that this is a theoretical lower bound designed to elucidate general findings and is impossible to calculate generally in this manner if given a specific region where markers all have different allele frequencies.

In regions of no intermarker LD between pairs of true markers ($r^2_{j,k} = 1$; $j,k < m$; $j \neq k$), probabilities of success are independent, and, under the assumption of positive dependence between tests of association at successive markers, the upper bound of $\Pr[R(i)]$ is a function of the power at each locus:

$$B_u \equiv \Pr[R(i)] \leqslant \Pr[I_{P_1(i)} = 1]\left\{1 - \prod_{j=1}^{M} \Pr[I_{P_2(j)} = 0]\right\} .$$

In this situation, we make a standard Bonferroni correction ($\alpha' = \alpha/KM$). Note that, when $M = m = 1$, $B_l = B_u$, and figure 2 shows these exact probabilities in regions of a single representative marker as red lines. See appendix A for a discussion of the rationale for the conditions required to achieve upper and lower bounds of $\Pr[R(i)]$.

"Replication" of Markers Not Detected in Initial Study

Let $\bar{R}(i)$ be the event that there is a local replication but that the representative marker is not successful in the second test. For $M > 1$,

$$\Pr[\bar{R}(i)] = \Pr[I_{P_1(i)} = 1] \Pr[I_{P_2(i)} = 0 \text{ and at least one of } I_{P_2(j)} = 1, j \neq i] .$$

Note that, by definition, $\Pr[\bar{R}(i)] = 0$ when $M = 1$. In regions of no intermarker LD, the maximum probability of $\bar{R}(i)$ can be determined:

$$\Pr[\bar{R}(i)] \leqslant \Pr[I_{P_2(i)} = 1]\left\{\Pr[I_{P_2(i)} = 0] - \prod_{j=1}^{M} \Pr[I_{P_2(j)} = 0]\right\} .$$

In regions of complete intermarker LD, $\Pr[\bar{R}(i)] = 0$.

Correction for Multiple Testing

For an overall type I error of α, we correct for multiple testing by setting the Bonferroni-corrected significance rate, α', as a function of a nominal level, the total number of markers tested, and the LD structure among them. All markers in the original study are assumed to be independent; so, for an original study involving M markers, $\alpha' = \alpha/M$, ensuring an overall type I error rate of α. Suppose a replication study involves the examination of K independent regions. Under our parameterization of the problem, we consider each region to be replicated independently. Thus, we require a difference in the correction for multiple testing to be applied within the region, compared with between regions, and so we have assumed that each region will have its own overall type I error rate of $\alpha' = \alpha/K$. Within each region, the error rate is further corrected according to the effective number of independent tests performed within this region. Hence, for region k, consisting of M^{I}_K independent and M^{D}_K dependent markers, $\alpha' =$

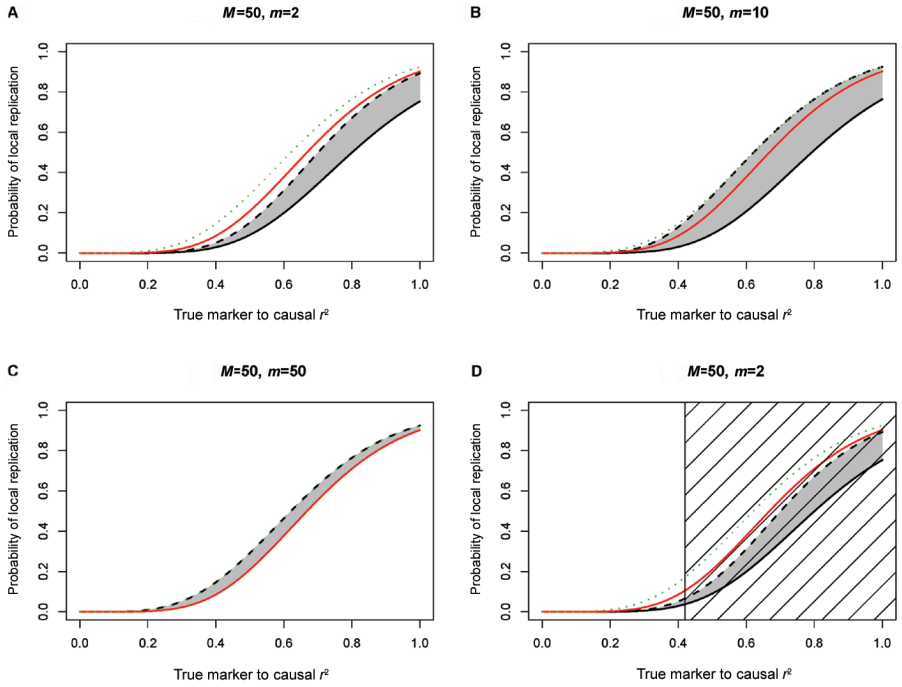

Figure 2. Theoretical probabilities of observing local replication in a region of 50 markers ($M = 50$) as a function of the common value of LD between the true marker and causal alleles. The number, *m,* of true markers is shown above the panels. The dotted green line shows the probability of achieving a significant result at the representative marker in the original study only. The red line shows the probability of achieving an exact replication. Since original and replicate study samples are assumed to be independent, the value at the red line is the product of the value at the green line and the probability of achieving a significant result at the representative marker in a replicate study when no additional markers are tested. The gray-shaded region represents the range of replication probabilities at all possible levels of LD between marker and disease loci. The upper bound of the gray-shaded region represents the maximum probability of replication for any given level of LD between each marker and a causal locus (*X*-axis), which occurs when all markers in the region are independent. This upper bound is represented with a dashed black line to emphasize the fact that, when multiple markers in a given region are independent, the possible value of LD between each marker and a single causal locus is constrained, and so the upper bound cannot be attained at all levels of LD between the marker and causal loci. The lower bound of the gray-shaded region represents the minimum probability of replication for any given level of LD between each marker and a causal locus, which occurs when all markers in the region are independent. This lower bound can be attained for all levels of LD between the marker and causal loci. The disease prevalence is 0.05, the GRR is 1.3, and the frequency of the high-risk allele is 0.25. In the first stage, 500,000 markers are genotyped for 3,000 cases and 3,000 controls. In the replication study, $K = 10$ regions are identified, and markers are genotyped for 1,500 cases and 1,500 controls. To ensure an overall type I error rate of 0.05 in the replicate study, the Bonferroni corrected rate is then $\alpha' = 0.05/[10 \times (50 - m + 1)]$ when the true markers are dependent and is $\alpha' = 0.05/(10 \times 50)$ when the true markers are independent. Panel D duplicates panel A to highlight the specific region for this example in which the maximum probability of replication cannot be attained, emphasizing the implicit requirement of allelic heterogeneity (multiple disease alleles) when the r^2 between marker and causal alleles is sufficiently large. For these sample parameter values, this occurs for $r^2 > 0.42$. (See appendix A for details on calculation of this cutoff value.)

$\alpha/K(M_K^1 + 1^{M_K^0})$. The upper and lower bounds of replication probabilities are found at these extreme conditions of complete and absent intermarker LD. Our correction is highly conservative in regions of no intermarker LD and allows for the detailed examination of replication within a set of independent regions.

Examining Variation across Genes ATGL161 *and* IL23R

To provide an example, to illustrate the effect of real LD patterns, we use data from the International HapMap Project[31] to compare the probability of local and exact replication in a simulation study designed to mimic primary and replicate studies that aim to find associations at markers in LD with SNP *rs2241880* (minor-allele frequency [MAF] of 0.453) in the *ATG16L1* gene and SNP *rs7517847* (MAF = 0.403) in the *IL23R* gene (MIM 607562). Both SNPs are known to be associated with Crohn disease (MIM 266600).[7,32] Using genotype data from CEPH samples (Utah residents with ancestry from northern and western Europe), we randomly select 50 markers within 100 kb of each SNP to represent additional markers selected for a replication study. We treat *rs2241880* and *rs7517847* as causal SNPs, and, for each randomly selected additional marker, we calculate the LD between the marker and "causal" alleles. For each value of LD equal to 0.1, 0.2,..., 1 between the initially associated marker and the causal alleles, we then compute the maximum probability of local replication and the probability of exact replication for 10,000 simulations. The maximum probability of local replication occurs

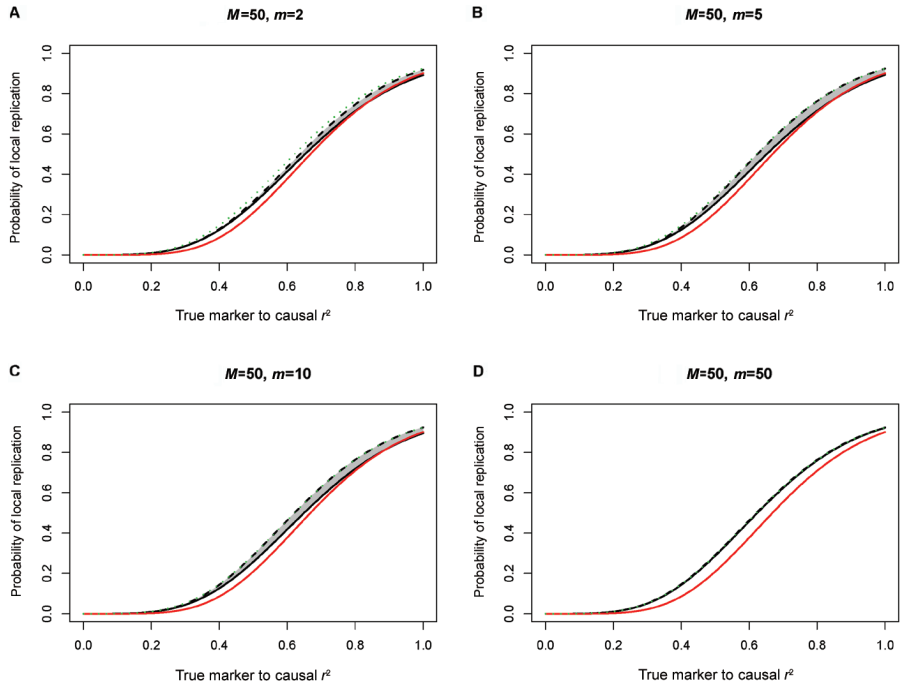

Figure 3. Theoretical probabilities of observing local replication in a region of $M = 50$ markers as a function of the common value of LD between $m - 1$ true marker and causal alleles. This graph is identical to figure 2 except that, instead of all m true markers having a common value of LD between marker and causal alleles, one of the additional markers selected is independent of all other markers and is in perfect LD with the causal variant. See the figure 2 legend for more details on the graphs. Values for m are shown above the panels.

when markers are independent; clearly, the markers are not independent here, and so these estimates of local replication probability are biased upward and thus present the best possible case for local replication.

Disease prevalence, genotype relative risk (GRR), the frequency of the high-risk allele, and all other study parameters are as described in the figure 2 legend. Since our purpose here is to compare the relative probabilities of exact and local replication, the exact values of these study parameters do not affect our findings.

Results

Power of Replication under Exact and Local Strategies

To explore formally the different contexts of local and exact replication, we compare the power to achieve each under two strategies, using theoretical expectations under the limiting conditions of complete and absent intermarker LD (see the "Methods" section). The results of these assessments elucidate several trends.

The shaded region in figure 2*A*–2*D* represents the probabilities of local replication across the range of intermarker LD as a function of the LD between the true marker and causal alleles; the lower boundary in each panel represents complete intermarker LD, and the upper boundary represents absent intermarker LD. The red line represents the probability of achieving an exact replication. Notice that, in all panels, the probability of replication at the lower boundary is always less than the probability of exact replication, illustrating the fact that, at the extreme of complete intermarker LD in a region, the local strategy cannot improve on the exact strategy. This is because the addition of any true markers in such a region is, of course, entirely redundant, and the inclusion of false markers will always reduce the probability of local replication in any region, as a simple consequence of the resulting penalties required to correct for multiple testing.

As intermarker LD decreases, the probability of local-replication success with a local strategy increases. The extent to which the local strategy is beneficial is strongly influenced by the proportion of genuinely associated markers among those added in the replication study (we refer to markers that are genuinely associated with the disease loci as "true"; see the "Methods" section for our operational definition). Figure 2*A*, 2*B*, and 2*C* represents situations with 2/50, 10/50, and 50/50 true markers, respectively. When very few of the new markers are true, the exact strategy performs better than the local strategy (fig. 2*A*). As the proportion of true markers increases, the local strategy becomes increasingly effective, as can be seen by the upper bound of the local replication strategy exceeding the power of the exact strategy in figure 2*B* and 2*C*. This is unsurprising because the addition of new markers in a region that has many true but initially uncaptured loci would be expected to increase power. At the same

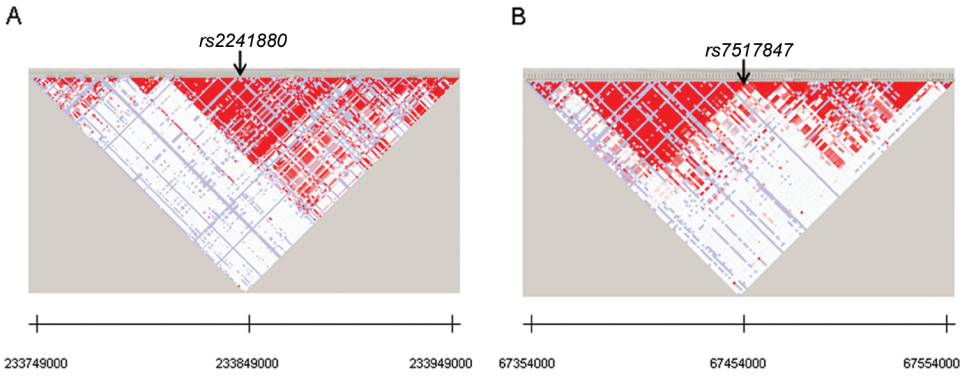

Figure 4. Pairwise r^2 plot for the HapMap CEU data from release 22, April 2007. The intensity of the shading is proportional to the value of r^2. *A, rs2241008,* at 233,848,107 bp on chromosome 2, contained within a single block of LD. *B, rs7517847,* at 67,454,257 bp on chromosome 1, between two blocks of LD.

time, however, it is increasingly penalized by the corrections required for multiple testing. When many of the new markers are not associated with the disease locus, as is the case in 48 of the 50 markers in figure 2*A*, the penalty becomes so severe that it offsets the increased power.

Examples illustrated in figure 2 assume that all true markers tested in the replication study have a common value of LD between the marker and causal alleles. The best case for local replication occurs when one or more of the additional true markers selected for the replication study is in greater LD with the causal variant than is the marker identified in the primary study. Figure 3 contrasts figure 2, showing changes to the upper and lower bounds of the probability of local replication in the extreme situation where one of the additional markers selected is independent of all other markers and is in perfect LD with the causal variant. The upper and lower bounds of the probability of local replication have increased and have converged, and, with a sufficient number of true markers, the probability of local replication is greater than that of exact replication at all values of LD between the remaining additional marker and causal alleles. In general, the strength and number of true markers, as well as the ratio of true to false markers added in a local replication design, cannot be known in advance, so it is practically impossible to predict the optimal selection of markers in advance of a replication study.

In a simulation of possible outcomes for replication of associations across genes known to be associated with Crohn disease, we considered 100-kb regions around SNPs *rs2241880* and *rs7517847*. *rs2241880* lies in a region of high intermarker LD; *rs7517847* lies in a region of medium-to-low intermarker LD, as shown in figure 4. Figure 5*A* shows that, when a region of high intermarker LD is tested, as the probability of exact replication decreases, the maximum probability of local replication improves relative to the probability of exact replication. This is because the local strategy provides a good chance of picking up a marker in higher LD with the causal variant than the initially identified variant when the initial finding is weak. Figure 5*B* confirms that, when a region of low intermarker LD is tested, the exact strategy is generally optimal when the initial finding is strongly associated with disease and that the local strategy is increasingly dependent on additional marker selection as the strength of the initial finding decreases.

Overall, the benefits of the local strategy are most prominent in the presence of allelic heterogeneity and low LD. In regions of incomplete intermarker LD, the level of LD between any marker and a causal locus is constrained.[29] Therefore, when only a single disease locus is present in a low-LD region, it is not possible for multiple markers to each be highly correlated with it. The results in figure 2*D* expand those of figure 2*A* to show that, under that particular scenario, if there is only a single disease locus, then the upper bound of the local replication power is only feasible when the marker-disease r^2 is less than a given value (for the particular parameter values of our example, $r^2 <\sim 0.42$) determined by constraints on the haplotype frequencies and the r^2 values imposed by the local LD patterns (appendix B). This constraint will become more severe as the number of added markers increases, thereby further lowering the permissible LD levels between the marker and disease locus. Unless there is allelic heterogeneity, the selection of appropriate markers required to optimize a strategy of local replication in a region of low intermarker LD may be too challenging in practice to yield success.

As a by-product of these considerations, it is important to emphasize the obvious point that the power of the initial GWA study is paramount in determining the probability of replication—that is, the probability of initially detecting a locus and then replicating it in an independent sample cannot exceed the power of the initial study. This is important to recognize in the context of GWA studies because a number of strategies aim to reduce genotyping

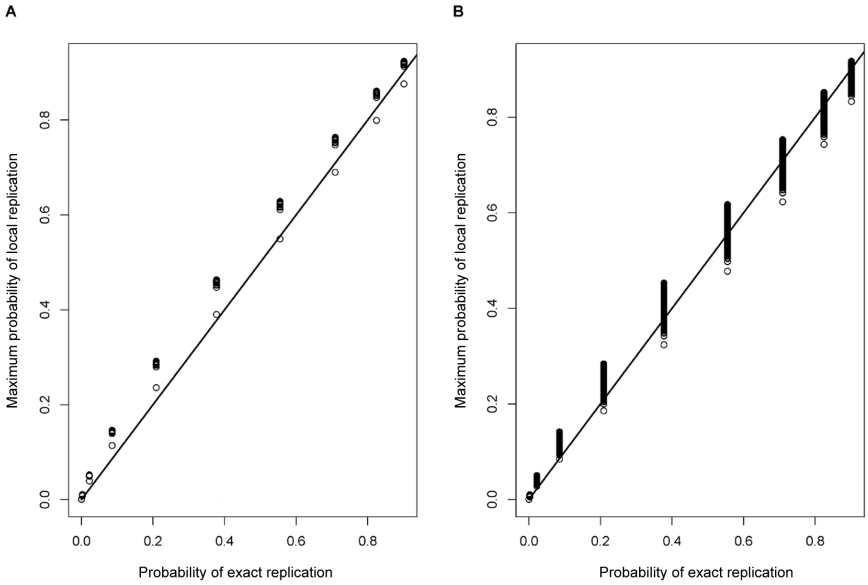

Figure 5. Maximum probability of local replication as a function of the probability of exact replication in simulations designed to mimic the outcomes of replication that attempts to find an association under the assumption that *rs2241880* (*A*) and *rs7517847* (*B*) are causal SNPs. Each point corresponds to a single simulation. The solid black line is for reference only, indicating when exact and maximum local replication probabilities are equal. See the "Methods" section for full details on simulations.

costs by minimizing the genotyping in the initial GWA study, at the cost of the power of initial detection.[27]

The example illustrated in figure 2 involves a replication study with the same type I error but with lower power than that of the original study. This example was selected to provide optimal visual clarification of the possible differences between the exact and local approachs. In other examples, the value of power and the significance thresholds used in the primary and replication studies will alter the absolute value of the results but will have no bearing on the relative merits of replication under the exact strategy versus under the local strategy.

In summary, our findings indicate that the effectiveness of the local strategy increases with the number and strength of true markers among the additional markers included in the replicate study. Results suggest that, when the original marker is strongly associated with disease—either because there is a large effect or because it is highly correlated with the causal variant—then an exact strategy is the best approach. This is because, in regions of low intermarker LD where the local strategy performs best, the chance of finding one or more additional markers that have a higher correlation with the causal variant than the originally identified variant does is small, and, in regions of high intermarker LD where there are perhaps additional highly correlated markers, the differences between the local and exact strategies are minimal. However, when the original marker is only weakly associated, the local strategy is likely to be a good approach in regions of high intermarker LD, because there is good chance of picking up at least one marker that is in greater LD with the causal variant than is the original marker. Although it is theoretically possible to benefit from a local replication strategy for a weakly associated initial finding in a region of low intermarker LD, it depends critically on the selection of true markers. If cost is an issue, then the exact approach is generally the better strategy, because even when additional markers do improve the effectiveness of the local strategy, the difference between the maximum probability of local replication and that of exact replication is marginal.

Fine-scale assessments of other markers in the region, although perhaps justified on biological and genetic grounds, will usually be more fruitfully pursued in samples other than those used to validate the initial results. Fine-scale mapping studies of the initial GWA samples may be useful for identifying other correlated markers for subsequent follow-up, but it is important to recognize the potential biases in initial effect-size estimates that will accompany such a strategy (i.e., the "winner's curse").[33]

"Replicating" Markers Other Than Those Initially Identified

Several studies have reported associations with markers other than those initially reported, even when the initial markers were tested among those added in the follow-up study,[28] with the differences often attributed to allelic or genetic heterogeneity. Figure 6 shows that this event can occur frequently without allelic heterogeneity.

At the extreme of complete intermarker LD, the probability of successful replication with some marker other than tthe original representative locus is zero, since highly

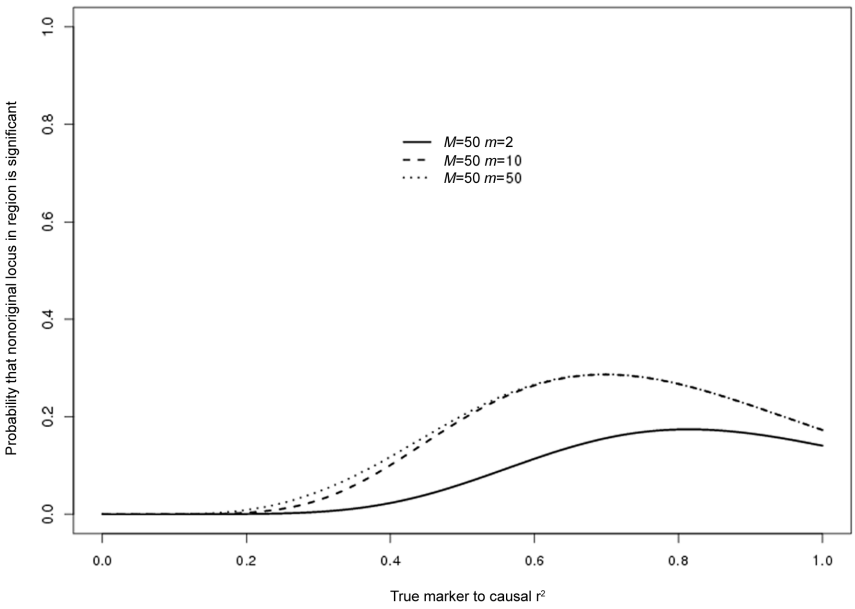

Figure 6. Maximum probability that a locus exceeds the significance threshold in a replication study but is different from the locus initially identified. Each line shows results for a different number of true markers (m), as indicated, tested in a region of $M = 50$ markers in the replicate study and corresponds to the probability that a single true marker is identified in the first study but not in the second study, even though the second study has sufficient power to detect similar nearby markers. The disease prevalence, risk, allele frequencies, sample sizes, marker designs, and type I error are as described in the figure 2 legend.

dependent markers cannot be simultaneously successful and unsuccessful. In regions of incomplete LD and/or for weak power in the replication study, the chances of detecting an association with some locus in a region increases with the number of new variants tested, at a rate that depends on the proportion of genuinely associated markers, whereas the probability of replicating the original representative marker decreases slowly, as a result of the increasingly conservative correction required. In this situation, the use of the local replication approach appears to be useful as an extension of association discovery, although not as a validation tool, since such results would initiate the need for further validation by another mechanism.

Discussion

We have examined two strategies commonly employed in replication studies of genetic association: the exact strategy, in which only significant loci are subsequently tested in a replication study, and the local stategy, in which additional variants are also included. We found that, in general, the exact strategy is more balanced in power and efficiency, in terms of cost and ability to replicate a maximum number of loci. However, it is possible to benefit from testing additional markers, especially when a region of high intermarker LD is tested and the initial finding is weakly associated with disease. This might occur, for example, when coverage in the initial GWA is poor. It is important to note that, when the local strategy does improve power, the patterns of association can often reveal different loci than those initially identified, thus rendering the interpretation difficult and likely requiring yet another validation study.

A precise value for local replication probability or tighter bounds on the range of local replication probabilities would require accurate consideration of the LD structure between markers and causal variants and could not be used for general inference. The local replication strategy also places strong demands on accurately accounting for the dependence between multiple tests. We have used the boundary conditions of complete or zero LD to obtain the range of possible values in figure 2. At these boundaries, multiple-testing corrections are straightforward. If the multiple-testing procedure used in any specific study is anticonservative, then local replication will inflate false-positive rates for apparent replication, whereas, if it is conservative, then true replications will be missed. Permutation procedures or nonfrequentist approaches to follow-up association validation may offer more flexibility and utility to local replication; Bayesian methods are particularly beneficial when markers are selected with prior information based on a previously significant association or functional relevance.[33]

For practical purposes and without loss of generality, we have assumed that significant markers to be replicated can be separated into regions. This is the basis for defining K regions to be replicated and for defining local replication

success on a region-by-region basis. If, instead, there are multiple significant markers in a small region, then this may serve to increase the probability of local replication for any one of these markers, but comparisons between exact and local strategies for a particular marker proceed as usual. Marker density has an effect here implicitly, in terms of the intermarker LD, which is modeled only at the extremes for multiple markers in any given region. Intermediate values of LD and associated marker density are not considered explicitly. This is an appealing aspect of the theoretical method we have devised for comparisons of exact and local strategies and makes it applicable to a wide range of factors that influence the performance of each strategy.

Many ongoing GWA studies are multistage in design. In general, these are not as powerful as replication strategies per se, but, instead, are used for efficient primary detection of new loci in large samples.[27] Our focus here was not on such studies but was on the next step in the process of disease-gene characterization: verification of the initial GWA findings in independent samples. In this regard, our results call for a clear distinction between fine mapping and replication. Novel loci detected by GWA are candidates for replication. Conversely, when a disease gene or region has already been confirmed, fine-mapping studies may help clarify the specific variants involved. The exact strategy is clearly a replication approach, one which has the added benefit of offering the possibility of further combining replication data with those from the initial GWA scan. The local strategy, however, combines both replication and fine mapping and, in doing so, increases ambiguity in the outcomes.

Acknowledgment

We thank David E. Evans for helpful discussions and comments on the manuscript.

Appendix A

Bounds for the Joint Probability Distribution of Standard Tests of Association at Multiple Markers

An upper bound for the probability that all tests are unsuccessful in a replicate study is given by

$$\Pr[S_2(M) = 0] < \max\{\Pr[I_{P_2(1)} = 0], \Pr[I_{P_2(2)} = 0], \ldots, \Pr[I_{P_2(M)=0}]\} .$$

Note that the upper bound occurs naturally when markers are in complete LD, so that the tests are statistically dependent, in which case

$$\Pr[S_2(M) = 0] = \Pr[I_{P_2(1)} = 0] . \quad \text{(A1)}$$

Under the assumption of positive dependence between successive tests at all M markers in the replication study,[34] a lower bound is given by

$$\Pr[S_2(M) = 0] > \Pr[I_{P_2(1)} = 0]\Pr[I_{P_2(2)} = 0], \ldots, \Pr[I_{P_2(M)} = 0] .$$

If successive standard tests of association at all M markers in a replication study are independent, then the probability that all tests are unsuccessful is the product of the marginal probabilities that each test is unsuccessful:

$$\Pr[S_2(M) = 0] = \Pr[I_{P_2(1)} = 0]\Pr[I_{P_2(2)} = 0], \ldots, \Pr[I_{P_2(M)} = 0] . \quad \text{(A2)}$$

That is, the lower bound occurs when markers are in complete linkage equilibrium, so that the tests are statistically independent.

Application of equations (A1) and (A2) to equation (1) gives lower and upper bounds, respectively, for $\Pr[R(i)]$ in the case of positive dependence between tests of association at successive markers. When there is no positive dependence between tests at successive markers, a nonzero lower bound cannot be found for the joint distribution. Consideration of equation (1) indicates that, at this extreme, the probability of local replication reduces to the probability of success in the first test.

Appendix B

Restrictions on Intermarker LD between Two Marker Loci as a Result of LD between the Marker and Causal Alleles

Consider three biallelic loci, labeled *A, B,* and *C,* with major alleles *A, B,* and *C* and minor alleles *a, b,* and *c,* respectively. Let the frequency of allele *A* be p_A. Let the frequency of allele *B* on chromosomes carrying an *A* (*a*) allele be $p_{B|A}$ ($p_{B|a}$). Then, the frequency of allele *B* is $p_B = p_{B|A}p_A + p_{B|a}(1 - p_A)$, and the LD between loci *A* and *B* can be expressed as

$$r^2_{AB} = (p_{B|A} - p_{B|a})^2 \times \frac{p_A(1 - p_A)}{p_B(1 - p_B)} . \quad \text{(B1)}$$

Similarly, let the frequency of allele *C* on chromosomes carrying an *A* (*a*) allele be $p_{C|A}$ ($p_{C|a}$). Similar results then hold for the frequency of allele *C*, p_C, and for the LD between loci *A* and *C*, r^2_{AC}.

Suppose that the joint distribution between loci *A* and *B* and that between loci *A* and *C* are identical. Specifically, let $p_{B|A} = p_{C|A}$ and $p_{B|A} = p_{C|a}$; then, for example, $r^2_{AB} = r^2_{AC}$, $p_{AC} = p_{AB}$, $p_{aC} = p_{aB}$, $p_B = p_C$, and

$$r^2_{BC} = \frac{(p_{BC} - p_B^2)^2}{p_B^2(1 - p_B)^2} . \quad \text{(B2)}$$

Using the fact that the sum of the various joint distri-

butions of alleles at the three loci must equal the appropriate joint distribution of alleles at any two of the loci, we can derive the constraints

$$\max(0, p_{AB} - p_{Ab}) \leq p_{ABC} \leq p_{AB} \tag{B3}$$

and

$$\max(0, p_{aB} - p_{ab}) \leq p_{aBC} \leq p_{aB} \ . \tag{B4}$$

If we let p_{ABC} and p_{aBC} take on their maximum values—p_{AB} and p_{aB}, respectively—then p_{BC}: $= p_{ABC} + p_{aBC} = p_{AB} + p_{aB} = p_B$, and, from expression (B2), it is easy to see that $r^2_{BC} = 1$. Hence, for any given common value of LD between loci *A* and *B* and between loci *A* and *C,* the LD between loci *B* and *C* can always take on the maximum value 1.

From expression (B2), it can also be seen that $r^2_{BC} = 0$ if and only if $p_{BC} = p^2_B$. Since $p_{BC} = p_{ABC} + p_{aBC}$, summation of the constraints (B3) and (B4) gives the following constraint on p_{BC}:

$$\max(0, p_{AB} - p_{Ab}) + \max(0, p_{aB} - p_{ab}) \leq p_{BC} \leq p_B \ .$$

Hence, if p_B can be selected such that $p^2_B < S$, where

$$\begin{aligned} S &= \max(0, p_{AB} - p_{Ab}) + \max(0, p_{aB} - p_{ab}) \\ &= \max[0, p_A(2p_{B|A} - 1)] + \max[0, (1 - p_A)(2p_{B|a} - 1)] \ , \end{aligned}$$

then it will not be possible to have $p_{BC} = p^2_B$, and so $r^2_{BC} > 0$. There are three cases:

(i) When $p_{B|a} < 0.5$ and $p_{B|A} < 0.5$, then $S = 0$.
(ii) When $p_{B|a} \geq 0.5$ and $p_{B|A} \geq 0.5$, then $S = 2p_B - 1$.
(ii) When $p_{B|a} < 0.5$ and $p_{B|A} \geq 0.5$ or when $p_{B|A} < 0.5$ and $p_{B|a} \geq 0.5$.

In cases (i) and (ii), p_B would have to be negative in order that $p^2_B < S$; thus, r^2_{BC} is unrestricted here. Also observe that, by considering the maximum and minimum values of $p_{B|a}$ and $p_{B|A}$ in these two cases, expression (B1) indicates that $0 \leq r^2_{AB} < p_A/(1 + p_A)$. In case (iii), expression (B1) indicates that $r^2_{AB} \geq p_A/(1 + p_A)$, and it is possible to select p_B so that $p^2_B < B$. Hence, r^2_{BC} can be restricted here, and the restriction gets increasingly severe as $|p_{B|a} - p_{B|A}|$ approaches 1 or, equivalently, as r^2_{AB} approaches 1. Thus, if the common value of LD between loci *A* and *B,* r^2_{AB}, or between loci *A* and *C,* r^2_{AC}, is $> p_A/(1 + p_A)$, then the LD between loci *B* and *C,* r^2_{BC}, is constrained to be nonzero. Conversely, as r^2_{BC} decreases to zero, the maximum value of r^2_{AB} decreases to $p_A/(1 + p_A)$.

Web Resource

The URL for data presented herein is as follows:

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for *IL23R* and Crohn disease)

References

1. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108
2. Thomas DC, Haile RW, Duggan D (2005) Recent developments in genomewide association scans: a workshop summary and review. Am J Hum Genet 77:337–345
3. Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109–118
4. Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429:446–452
5. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, et al (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37: 1243–1246
6. Davey Smith G, Ebrahim S, Lewis S, Hansell AL, Palmer LJ, Burton PR (2005) Genetic epidemiology and public health: hope, hype, and future prospects. Lancet 366:1484–1498
7. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A, et al (2006) A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. Science 314:1461–1463
8. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, et al (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447:1087–1093
9. Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A, et al (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 39:631–637
10. Haiman CA, Patterson N, Freedman ML, Myers SR, Pike MC, Waliszewska A, Neubauer J, Tandon A, Schirmer C, McDonald GJ, et al (2007) Multiple regions within 8q24 independently affect risk for prostate cancer. Nat Genet 39:638–644
11. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, Jonasdottir A, Sigurdsson A, Baker A, Palsson A, et al (2007) A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 316: 1491–1493
12. Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, Illig T, Wichmann HE, Meitinger T, Hunter D, Hu FB, et al (2006) A common genetic variant is associated with adult and childhood obesity. Science 312:279–283
13. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, et al (2007) A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39:870–874
14. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, et al (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389
15. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33:177–182
16. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin

P, Vincent D, Belisle A, Hadjadj S, et al (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881–885
17. Smyth LJ, Elkord E, Taher TE, Jiang HR, Burt DJ, Clayton A, van Veelen PA, de Ru A, Ossendorp F, Melief CJ, et al (2006) CD8 T-cell recognition of human 5T4 oncofetal antigen. Int J Cancer 119:1638–1647
18. The 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
19. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, et al (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 39:645–649
20. Colhoun HM, McKeigue PM, Davey Smith G (2003) Problems of reporting genetic associations with complex outcomes. Lancet 361:865–872
21. Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2:e124
22. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nat Genet 29:306–309
23. Ober C (2005) Perspectives on the past decade of asthma genetics. J Allergy Clin Immunol 116:274–278
24. Williams NM, Preece A, Morris DW, Spurlock G, Bray NJ, Stephens M, Norton N, Williams H, Clement M, Dwyer S, et al (2004) Identification in 2 independent samples of a novel schizophrenia risk haplotype of the dystrobrevin binding protein gene (*DTNBP1*). Arch Gen Psychiatry 61:336–344
25. Palmer LJ, Cardon LR (2005) Shaking the tree: mapping complex disease genes with linkage disequilibrium. Lancet 366: 1223–1234
26. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, et al (2007) Replicating genotype-phenotype associations. Nature 447:655–660
27. Neale BM, Sham PC (2004) The future of association studies: gene-based analysis and replication. Am J Hum Genet 75: 353–362
28. Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213
29. Terwilliger JD, Hiekkalinna T (2006) An utter refutation of the "fundamental theorem of the HapMap." Eur J Hum Genet 14:426–437
30. Bacanu SA, Devlin B, Roeder K (2000) The power of genomic control. Am J Hum Genet 66:1933–1944
31. The International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796
32. Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, Albrecht M, Mayr G, De La Vega FM, Briggs J, et al (2007) A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in *ATG16L1*. Nat Genet 39:207–211
33. Greenland S (2007) Bayesian perspectives for epidemiological research. II. Regression analysis. Int J Epidemiol 36:195–202
34. Gleser LJ, Moore DS (1985) The effect of positive dependence on chi-squared tests for categorical data. J R Stat Soc Ser B (Methodological) 47:459–465