nature genetics

Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies

Andrew D Skol, Laura J Scott, Gonçalo R Abecasis & Michael Boehnke

Genome-wide association is a promising approach to identify common genetic variants that predispose to human disease[1–4]. Because of the high cost of genotyping hundreds of thousands of markers on thousands of subjects, genome-wide association studies often follow a staged design in which a proportion ($\pi_{samples}$) of the available samples are genotyped on a large number of markers in stage 1, and a proportion ($\pi_{samples}$) of these markers are later followed up by genotyping them on the remaining samples in stage 2. The standard strategy for analyzing such two-stage data is to view stage 2 as a replication study and focus on findings that reach statistical significance when stage 2 data are considered alone[2]. We demonstrate that the alternative strategy of jointly analyzing the data from both stages almost always results in increased power to detect genetic association, despite the need to use more stringent significance levels, even when effect sizes differ between the two stages. We recommend joint analysis for all two-stage genome-wide association studies, especially when a relatively large proportion of the samples are genotyped in stage 1 ($\pi_{samples} \geq 0.30$), and a relatively large proportion of markers are selected for follow-up in stage 2 ($\pi_{markers} \geq 0.01$).

Genome-wide association studies are now underway[5], enabled by rapidly decreasing genotyping costs, massively multiplexed genotyping technologies and the large-scale SNP discovery and genotyping efforts of the SNP Consortium[6], the HapMap project[7] and Perlegen Sciences[8]. These projects have identified and genotyped well over 1 million SNPs in several human populations, allowing investigators to select a set of genetic markers that efficiently assays most common human genetic variation[9–11]. Compared with one-stage designs that genotype all samples on all markers, well-constructed two-stage association designs maintain power while substantially reducing genotyping requirements[12–14].

The power of two-stage genome-wide association studies to identify variants that predispose to disease depends on a number of factors controlled by the investigator, including how markers are selected, how samples are divided between stages 1 and 2, the proportion of markers tested in stage 2 and the strategy used to test for association.

We focus on two-stage designs in which all M markers are genotyped in a proportion of the samples ($\pi_{samples}$) in stage 1, and results of stage 1 are used to select a proportion of these M markers ($\pi_{markers}$) for follow-up on the remaining samples in stage 2. These samples might be cases and controls for a genetic disease or individuals measured for a quantitative trait. We assume initially that the M markers are in linkage equilibrium.

Our purpose is to compare power for the standard replication-based analysis strategy with the power of the alternative strategy of joint analysis of all available samples. Both strategies can be tailored to achieve any desired genome-wide false positive rate (type I error rate) of α_{genome} so that the number of false positives expected in the genome-wide association scan is α_{genome}. In the replication strategy,

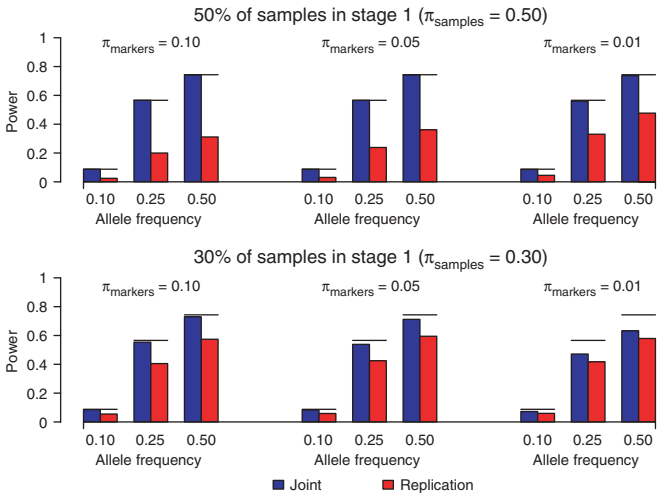

Figure 1 Power of a two-stage design for joint and replication-based analysis with 1,000 cases and 1,000 controls genotyped on 300,000 independent markers with $\alpha_{genome} = 0.05$. Uses a multiplicative genetic model with genotype relative risk (GRR) = P(caseIDD)/P(caseIDd) = P(caseIDd)/P(caseIdd) = 1.40 and prevalence of 0.10. The black line above each pair of bars indicates the power of the one-stage design in which all 2,000 samples are genotyped on all 300,000 markers.

Department of Biostatistics and Center for Statistical Genetics, University of Michigan, 1420 Washington Heights, Ann Arbor, Michigan 48109-2029, USA. Correspondence should be addressed to M.B. (boehnke@umich.edu).

Received 31 May 2005; accepted 5 November 2005; published online 15 January 2006; corrected after print 19 February 2006 (details online); doi:10.1038/ng1706

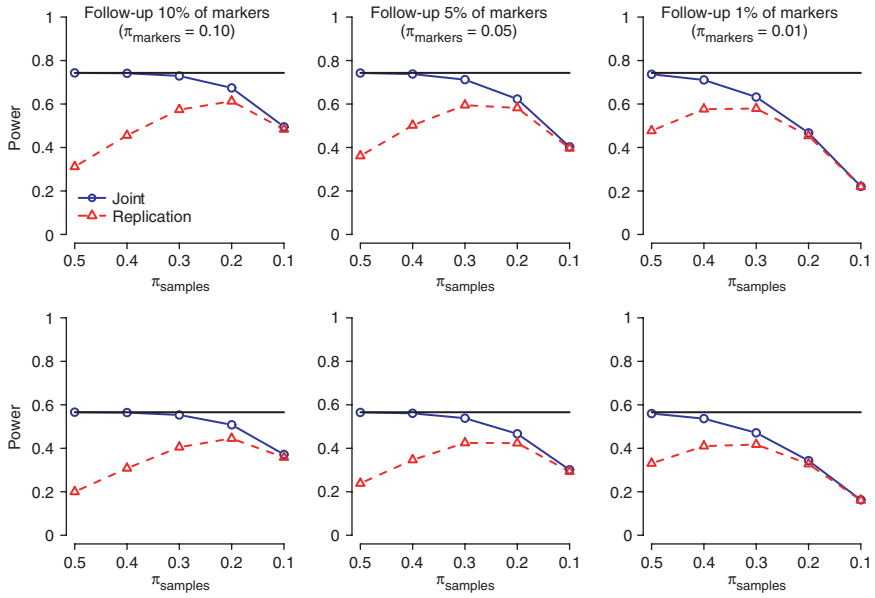

Figure 2 Power of a two-stage design for joint and replication-based analysis with 1,000 cases and 1,000 controls genotyped on 300,000 independent markers with $\alpha_{genome} = 0.05$, using a GRR of 1.40 and prevalence of 0.10. Control risk allele frequency is 0.50 for the upper row and 0.25 for the lower row. Horizontal black lines indicate the power of the one-stage design, in which all 2,000 samples are genotyped on all 300,000 markers.

genotype data from stage 2 samples are used to test for association using the Bonferroni-corrected significance level of $\alpha_{genome}/(\pi_{markers} \times M)$. In the joint analysis strategy, test statistics from stages 1 and 2 are combined, and a significance level of approximately α_{genome}/M is used. We show that joint analysis of the data almost always provides greater power than replication-based analysis of only stage 2 data, despite the more stringent significance level required by joint analysis. Replication-based analysis is preferable only when genetic effects are much larger in the stage 2 than in the stage 1 sample.

We compared the power of replication-based and joint analysis strategies for genome-wide association at $\alpha_{genome} = 0.05$ for a wide range of sample sizes, proportions of samples used in stage 1 ($\pi_{samples}$) and proportions of markers selected for follow-up in stage 2 ($\pi_{markers}$). We also examined multiple genetic models, effect sizes and frequencies of the variants predisposing to disease (see Methods). We determined the power of joint and replication-based analyses for six two-stage genome-wide association designs in which $\pi_{samples} = 50\%$ or 30% (**Fig. 1**). In these examples, joint analysis was often substantially more powerful than replication-based analysis, despite the need in joint analysis to use a more stringent significance level. For example, power for a two-stage design (**Fig. 1**) increased from 26% for replication-based analysis to 74% for joint analysis when 1,000 cases and 1,000 controls were split equally between the stages ($\pi_{samples} = 0.50$), 10% of stage 1 markers were used for follow-up in stage 2 ($\pi_{markers} = 0.10$), disease prevalence was 0.10, control allele frequency was 0.50, disease model was multiplicative and genotype relative risk (GRR) was 1.40.

Furthermore, the best strategy for analyzing two-stage genome-wide association data did not depend on proportion of samples used in stage 1 ($\pi_{samples}$) or the proportion of markers selected for follow-up in stage 2 ($\pi_{markers}$). Joint analysis was always more powerful than replication-based analysis in the examples displayed (**Fig. 2**) and often achieved power comparable to that of the more genotyping-intensive one-stage design. The advantage of joint analysis decreased as the proportion of samples $\pi_{samples}$ used in stage 1 decreased. For small values of $\pi_{samples}$, the power of joint and replication-based analysis strategies was comparable, because when $\pi_{samples}$ was small, the stage 1 information discarded by the replication-based analysis was modest. However, in that setting, variants that predisposed to disease were less likely to be selected for stage 2 follow-up, and both two-stage strategies typically had much lower power than the corresponding one-stage design.

The proportion of markers genotyped in stage 2 affected the power of joint and replication-based analyses in markedly different ways (**Fig. 3**). For joint analysis, as $\pi_{markers}$ decreased, the probability of selecting for follow-up a variant that predisposes to disease also decreased, resulting in less power. In

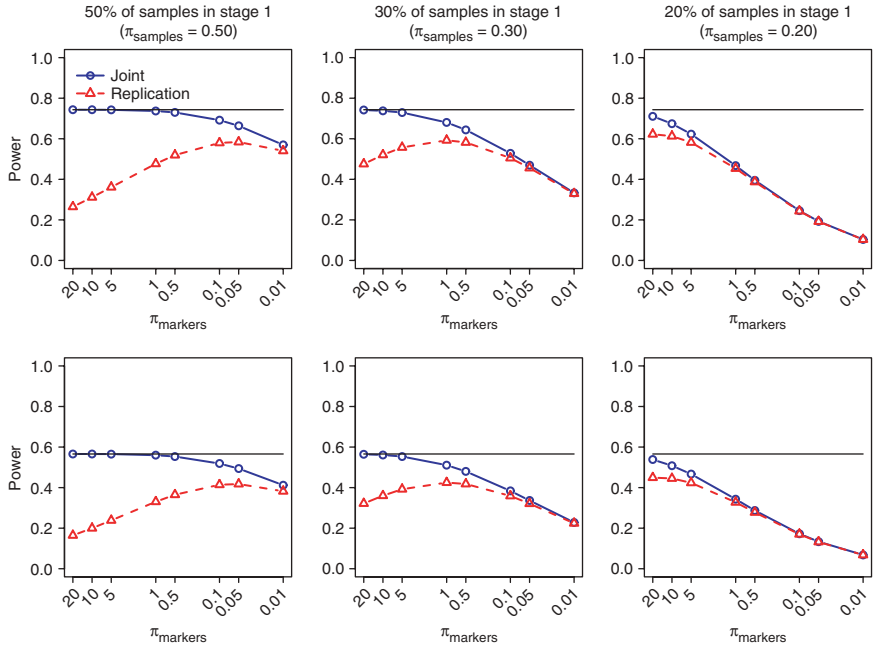

Figure 3 Power of a two-stage design for joint and replication-based analysis with 1,000 cases and 1,000 controls genotyped on 300,000 independent markers with $\alpha_{genome} = 0.05$, using a GRR of 1.40 and prevalence of 0.10. Control risk allele frequency is 0.50 for the upper row and 0.25 for the lower row. The black line indicates the power of the one-stage design in which all 2,000 samples are genotyped on all 300,000 markers.

Table 1 Significance thresholds and power of joint analysis for two-stage genome-wide association designs

						Power					
			Significance threshold			GRR = 1.30		GRR = 1.35		GRR = 1.40	
$\pi_{samples}$	$\pi_{markers}$	Proportion of genotypes[a]	C_1	C_2	C_{joint}	Joint	Rep	Joint	Rep	Joint	Rep
1.0	0	1.00	—	—	5.23	0.26	—	0.51	—	0.75	—
0.50	0.10	0.55	1.64	4.65	5.23	0.26	0.08	0.51	0.17	0.75	0.31
	0.05	0.53	1.96	4.50	5.23	0.26	0.09	0.51	0.21	0.75	0.36
	0.01	0.51	2.58	4.15	5.23	0.26	0.14	0.50	0.29	0.74	0.48
0.40	0.10	0.46	1.64	4.65	5.23	0.26	0.12	0.51	0.27	0.75	0.46
	0.05	0.43	1.96	4.50	5.23	0.26	0.14	0.50	0.30	0.74	0.51
	0.01	0.41	2.58	4.15	5.20	0.24	0.17	0.48	0.36	0.71	0.58
0.30	0.10	0.37	1.64	4.65	5.22	0.25	0.17	0.50	0.36	0.73	0.58
	0.05	0.34	1.96	4.50	5.21	0.24	0.18	0.48	0.37	0.72	0.60
	0.01	0.31	2.58	4.15	5.16	0.21	0.18	0.42	0.37	0.64	0.58
0.20	0.10	0.28	1.64	4.65	5.19	0.23	0.19	0.46	0.39	0.68	0.62
	0.05	0.24	1.96	4.50	5.16	0.21	0.18	0.42	0.38	0.63	0.59
	0.01	0.21	2.58	4.15	5.06	0.15	0.14	0.31	0.29	0.47	0.46

Shown is analysis of 1,000 cases and 1,000 controls, $M = 300{,}000$ markers, genome-wide significance level $\alpha_{genome} = 0.05$, multiplicative model, control risk allele frequency = 0.40 and prevalence = 0.10.
[a](Number of genotypes required for two-stage design) / (number of genotypes required when all markers are genotyped on all samples).

contrast, for replication-based analysis, power increased when fewer markers were selected for follow-up. This behavior is due to two competing effects: reducing $\pi_{markers}$ decreases the probability that a variant that predisposes to disease will be selected for genotyping in stage 2, but it increases the probability the variant will be found significant if it is genotyped in stage 2 because of the correspondingly less stringent significance threshold.

A two-stage design using joint analysis can achieve nearly the same power as the one-stage design in which all samples are genotyped on all markers (**Figs. 1–3**) but with much less genotyping (see also refs. 12–14 and **Table 1**). For example, a one-stage design that genotypes 1,000 cases and 1,000 controls on 300,000 markers (600 million genotypes) has 75% power to find a disease-predisposing variant with GRR = 1.40, prevalence 0.10 and risk allele frequency 0.50 in the controls. Nearly identical power (72%) can be achieved with only 34% as many genotypes by using $\pi_{samples} = 30\%$ in stage 1 and $\pi_{markers} = 5\%$ in stage 2. Note that for this sample size, both two-stage strategies examined here and even the corresponding one-stage design have low power to detect rare alleles with modest effects. Power calculations for arbitrary sample sizes and genetic models can be carried out using a tool available at our website (see Methods).

We compared joint and replication-based analysis for two-stage designs for a much broader set of genetic models, sample sizes and false positive rates. In every case, when we calibrated the two strategies to achieve the same genome-wide false positive rate (α_{genome}) we found that the joint analysis was more powerful (**Supplementary Tables 1** and **2** show additional examples). This makes sense, as joint analysis makes full use of stage 1 data, including the strength of evidence for the observed stage 1 association, whereas replication-based analysis uses only the information that the stage 1 association exceeds the threshold for follow-up but otherwise ignores the strength of the stage 1 evidence. By this same argument, joint analysis is more powerful in the presence of marker-marker linkage disequilibrium or multiple variants that predispose to disease.

Another level of complexity is added when heterogeneity in genetic effect size exists between samples used in stages 1 and 2. Such heterogeneity may arise for multiple reasons: for example, if investigators preferentially select cases from different geographic regions for each stage, or if cases for one stage have family histories of disease, whereas cases for the other stage do not. Unless the risk allele has a much larger effect in the stage 2 samples, joint analysis will remain more powerful than replication-based analysis (**Fig. 4**). Furthermore, note that because our analysis is based on combining test statistics rather than raw data across stages, it explicitly allows for heterogeneity between stages; if necessary, the statistics calculated for each stage can also be adjusted for within-stage heterogeneity[15–21].

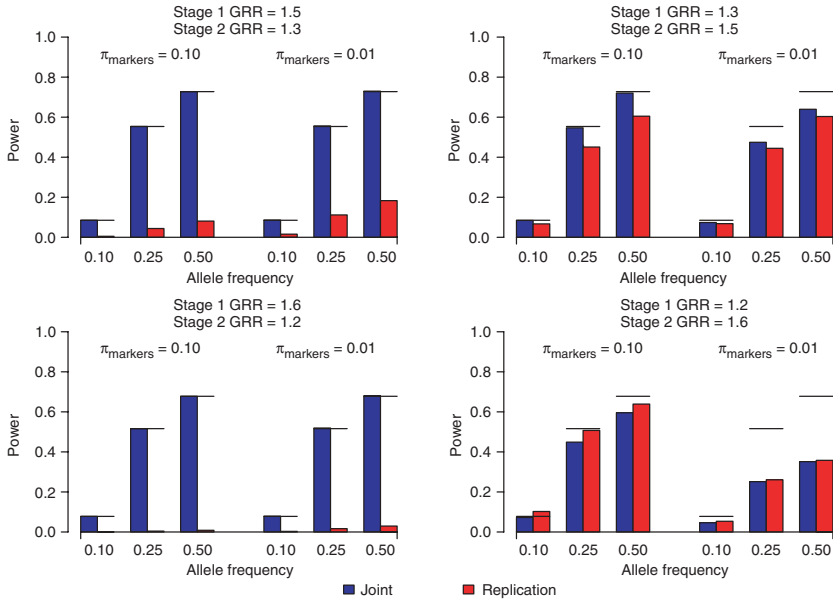

Figure 4 Power of a two-stage design for joint and replication-based analyses in the presence of between-stage heterogeneity with 1,000 cases and 1,000 controls genotyped on 300,000 independent markers with $\alpha_{genome} = 0.05$. The populations from which the stage 1 and 2 samples were drawn have prevalence 0.10 and control risk allele frequency of 0.10, 0.25 or 0.50, and the same number of samples are used for stages 1 and 2 ($\pi_{samples} = 0.50$). Graphs at left assume that the stage 1 samples have the higher genotype relative risk, and graphs at right assume that the stage 2 samples have higher genotype relative risk. The black line indicates the power of the one-stage design in which all 2,000 samples are genotyped on all 300,000 markers.

We repeated our power calculations, allowing for heterogeneity between stages, both in situations in which the effect size was stronger in stage 1 (**Fig. 4**, left) and in which the effect size was stronger in stage 2 (**Fig. 4**, right). Regardless of the difference between the genotype relative risks influencing the stage 1 and 2 samples, when the GRR is greater in stage 1, joint analysis is far more powerful than replication-based analysis. When the effect size is larger in stage 2, the replication-based analysis can sometimes be marginally more powerful than joint analysis. This can occur because replication-based analysis discards the data from stage 1, and although information is lost, greater power is achieved because the much stronger association in stage 2 samples is not diluted by the modest association in stage 1 samples.

For simplicity, we have ignored any consequence of difference in per-genotype cost for stage 1 and 2. Our investigation has focused on analysis strategy only, and regardless of differences in per-genotype costs for the two stages, joint analysis generally continues to be more powerful than replication-based analysis. However, the most cost-efficient allocation of samples to stages 1 and 2 will depend on differences in per-genotype costs at each stage.

Genome-wide association studies are already underway[5], and given the high cost and exciting potential of these studies, it is important that data be collected and analyzed efficiently. Two-stage and even multistage designs are being used because they can achieve nearly the same power as the one-stage design with substantially reduced genotyping. Regardless of the two-stage design used, analyzing the data from both stages jointly is almost always more powerful than treating stage 2 as a replication study for stage 1, and in many cases joint analysis results in substantially greater power. We suggest that for two-stage genome-wide association studies, we should forget replication and instead use joint analysis as the standard analysis strategy.

METHODS

We assume N cases and N controls are available for genotyping and that a proportion of these ($\pi_{samples}$) are genotyped in stage 1. Equivalent arguments hold for unequal numbers of cases and controls and other types of association studies, such as those based on a sample of individuals measured for a quantitative trait. Evidence for association at stage 1 is evaluated for each of the M markers and used to select approximately ($\pi_{markers} \times M$) markers for follow-up genotyping in the remaining ($(1 - \pi_{samples}) \times N$) cases and ($(1 - \pi_{samples}) \times N$) controls in stage 2. For simplicity, throughout the paper we refer to $\pi_{markers}$ as the proportion of markers genotyped in stage 2. In practice, though, we calculate power and critical values by requiring markers genotyped in stage 1 to have P values $< \pi_{markers}$, rather than selecting exactly $\pi_{markers} \times M$ markers for follow-up in each scan. The proportion of markers selected for follow-up will vary for each scan but will be very close to $\pi_{markers}$ when the number of disease-associated markers is small relative to $\pi_{markers} \times M$.

To evaluate evidence for association at stage 1, let $\hat{p}_1'$ and $\hat{p}_1$ be the estimated risk allele frequencies in cases and controls, respectively, and define the test statistic

$$z_1 = \frac{\hat{p}_1' - \hat{p}_1}{\sqrt{[\hat{p}_1'(1-\hat{p}_1') + \hat{p}_1(1-\hat{p}_1)]/(2N\pi_{samples})}}.$$

Under the null hypothesis of no association, and when a large number of samples ($N \times \pi_{samples}$) is genotyped in stage 1, z_1 follows a normal distribution with mean 0 and variance 1; quantiles for this standard normal distribution can be used to determine a threshold C_1 for selecting markers for follow-up such that $P(|z_1| > C_1) = \pi_{markers}$.

In a replication-based analysis, an analogous statistic z_2 is calculated using only stage 2 data and is compared with a new significance threshold C_2. To constitute a replication, we also require z_1 and z_2 to have the same sign, resulting in identification of the same risk allele. The false positive rate for a marker when using stage 1 and stage 2 significance thresholds of C_1 and C_2 is simply $\alpha_{marker} = P(|z_1| > C_1)\, P(|z_2| > C_2, \text{sign}(z_1) = \text{sign}(z_2))$.

In a joint analysis, a new statistic that allows for between-stage heterogeneity,

$$z_{joint} = \sqrt{\pi_{samples}}z_1 + \sqrt{1 - \pi_{samples}}z_2 \quad (1)$$

is compared with a significance threshold C_{joint}. As z_{joint} incorporates stage 1 genotype data, z_1 and z_{joint} are not independent even under the null hypothesis of no association. The false positive rate corresponding to thresholds C_1 and C_{joint} is $\alpha_{marker} = P(|z_1| > C_1 \text{ AND } |z_{joint}| > C_{joint}) = P(|z_1| > C_1)\, P(|z_{joint}| > C_{joint} \mid |z_1| > C_1)$, which can be calculated numerically by evaluating a simple integral (see equation (2) below). Achieving the same nominal significance level for replication and joint analysis generally requires quite different thresholds C_2 and C_{joint} (**Table 1** and **Supplementary Table 1**). Note that equation (1) allows evidence for association to be combined without assuming equal effect sizes and allele frequencies for the two stages. A statistic based on combining the raw data and assuming homogeneity between stages would result in slightly increased power in homogeneous samples at the risk of possibly inflated error rates or loss of power when there is heterogeneity between stages.

We next derive power estimates for the two analysis strategies. In all examples presented in this paper, we set $\alpha_{genome} = 0.05$ (in other words, the probability of observing at least one false positive result in the overall analysis of 300,000 markers is controlled to be 0.05). For simplicity, we assume that all markers are in linkage equilibrium so that the corresponding Bonferroni-corrected false positive rate for each marker $\alpha_{marker} = 0.05/M$. For a genome-wide scan with 300,000 markers, we set $\alpha_{marker} = \alpha_{genome}/300{,}000 = 1.67 \times 10^{-7}$. If the samples are split evenly between the two stages ($\pi_{samples} = 0.50$) and a false positive rate of 0.10 for stage 1 is used so that $\pi_{markers} = 0.10$, then $C_1 = 1.64$, $C_2 = 4.79$ and $C_{joint} = 5.23$. A software tool is available on our website (see below) to calculate C_1, C_2 and C_{joint} using any combination of number of markers M, number of samples N, $\pi_{samples}$, $\pi_{markers}$ and α_{genome}.

Stage 1. Power for stage 1 is the probability that a disease-predisposing variant is selected for follow-up in stage 2. To calculate power for stage 1, we describe the distribution of z_1 as a function of the risk allele frequencies p and p' for controls and cases. The statistic z_1 in large samples follows an approximate normal distribution with mean

$$\mu_1 = \frac{p' - p}{\sqrt{[p'(1-p') + p(1-p)]/(2N\pi_{samples})}}$$

and variance 1. Let $\Phi[x]$ be the cumulative distribution function for the standard normal distribution evaluated at x. Then, under the null hypothesis of no association, $\mu_1 = 0$ and $C_1 = \Phi^{-1}(1 - \pi_{markers}/2)$. The probability that a marker is selected for stage 2 genotyping is $P_1 = 1 - \Phi[C_1 - \mu_1] + \Phi[-C_1 - \mu_1]$.

Stage 2 for replication-based analysis. In a replication-based analysis, an analogous statistic is calculated using stage 2 genotype data only. The statistic z_2 in large samples follows an approximate normal distribution with mean

$$\mu_2 = \frac{p' - p}{\sqrt{[p'(1-p') + p(1-p)]/[2N(1 - \pi_{samples})]}}$$

and variance 1. Under the null hypothesis of no association, $\mu_2 = 0$ and $C_2 = \Phi^{-1}(1 - \alpha_{marker}/\pi_{markers})$, the critical value for the one-sided test of replication. The probability a variant that predisposes a disease will be significantly associated, given that it is selected for genotyping in stage 2, is

$$P_2 = (1 - \Phi[C_2 - \mu_2])\frac{1 - \Phi[C_1 - \mu_1]}{1 - \Phi[C_1 - \mu_1] + \Phi[-C_1 - \mu_1]} + \Phi[-C_2 - \mu_2]\frac{\Phi[-C_1 - \mu_1]}{1 - \Phi[C_1 - \mu_1] + \Phi[-C_1 - \mu_1]}$$

using a replication-based analysis. The power of the replication-based analysis is the product P_1P_2.

Stage 2 for joint analysis. Conditional on the observed stage 1 statistic $z_1 = a$, the statistic for joint analysis z_{joint} follows an approximate normal distribution in large samples with mean

$$\mu_{joint} = \frac{p' - p}{\sqrt{[p'(1-p') + p(1-p)]/(2N)}} + \sqrt{\pi_{samples}}(a - \mu_1)$$

and variance $(1 - \pi_{samples})$. Under the null hypothesis of no association, $\mu_{joint} = \sqrt{\pi_{samples}}a$. The critical value C_{joint} can be calculated iteratively by finding the threshold that satisfies $P(|z_{joint}| > C_{joint}|T) = \alpha_{marker}/(M\pi_{markers})$ under the null hypothesis, where T is the event $|z_1| > C_1$. The probability of detecting association in stage 2 in a joint analysis is

$$\begin{aligned} P_{joint} &= P(|z_{joint}| > C_{joint}|T) \\ &= \int_{-\infty}^{-C_1} [P(z_{joint} > C_{joint}|z_1 = x) + P(z_{joint} < -C_{joint}|z_1 = x)]f(x|T)dx \\ &+ \int_{C_1}^{\infty} [P(z_{joint} > C_{joint}|z_1 = x) + P(z_{joint} < -C_{joint}|z_1 = x)]f(x|T)dx \end{aligned} \quad (2)$$

where $f(x|T)$ is the probability density function z_1 given that $|z_1| > C_1$. This equation is also used to identify the critical value C_{joint} by allowing z_1 and z_{joint} to follow their null distributions. The power of the joint analysis is P_1P_{joint}.

URLs. Power calculations for arbitrary sample sizes and genetic models can be carried out using a tool available at our website (http://csg.sph.umich.edu).

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This research was supported by the US National Institutes of Health.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
2. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
3. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22**, 139–144 (1999).
4. Cardon, L.R. & Bell, J.I. Association study designs for complex diseases. *Nat. Rev. Genet.* **2**, 91–99 (2001).
5. Klein, R.J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
6. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
7. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
8. Hinds, D.A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
9. Johnson, G.C.L. *et al.* Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29**, 233–237 (2001).
10. Ke, X. & Cardon, L.R. Efficient selective screening of haplotype tag SNPs. *Bioinformatics* **19**, 287–288 (2003).
11. Stram, D.O. *et al.* Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the multiethnic cohort study. *Hum. Hered.* **55**, 27–36 (2003).
12. Satagopan, J.M., Venkatraman, E.S. & Begg, C.B. Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* **60**, 589–597 (2004).
13. Satagopan, J.M., Verbel, D.A., Venkatraman, E.S., Offit, K.E. & Begg, C.B. Two-stage designs for gene-disease association studies. *Biometrics* **58**, 163–170 (2002).
14. Thomas, D., Xie, R.R. & Gebregziabher, M. Two-stage sampling designs for gene association studies. *Genet. Epidemiol.* **27**, 401–414 (2004).
15. Devlin, B., Roeder, K. & Wasserman, L. Genomic control, a new approach to genetic-based association studies. *Theor. Popul. Biol.* **60**, 155–166 (2001).
16. Hinds, D.A. *et al.* Matching strategies for genetic association studies in structured populations. *Am. J. Hum. Genet.* **74**, 317–325 (2004).
17. Pritchard, J.K. & Donnelly, P. Case-control studies of association in structured or admixed populations. *Theor. Popul. Biol.* **60**, 227–237 (2001).
18. Ripatti, S., Pitkaniemi, J. & Sillanpaa, M.J. Joint modeling of genetic association and population stratification using latent class models. *Genet. Epidemiol.* **21**, S409–S414 (2001).
19. Satten, G.A., Flanders, W.D. & Yang, Q.H. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am. J. Hum. Genet.* **68**, 466–477 (2001).
20. Shmulewitz, D., Zhang, J.Y. & Greenberg, D.A. Case-control association studies in mixed populations: Correcting using genomic control. *Hum. Hered.* **58**, 145–153 (2004).
21. Yang, B.Z., Zhao, H.Y., Kranzler, H.R. & Gelernter, J. Practical population group assignment with selected informative markers: characteristics and properties of Bayesian clustering via STRUCTURE. *Genet. Epidemiol.* **28**, 302–312 (2005).

Corrigendum: Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies

Andrew D Skol, Laura J Scott, Gonçalo R Abecasis & Michael Boehnke

Nat. Genet. **38**, 209–213 (2006).

In Table 1 of the versions of this article initially published online and in print, the significance thresholds for C_2 were incorrect, and the significance thresholds for C_{joint} in the case of $\pi_{samples} = 0.20$ were incorrect. The error has been corrected in the HTML and PDF versions of the article. This correction has been appended to the PDF version.

Table 1 Significance thresholds and power of joint analysis for two-stage genome-wide association designs

						Power					
			Significance threshold			GRR = 1.30		GRR = 1.35		GRR = 1.40	
$\pi_{samples}$	$\pi_{markers}$	Proportion of genotypes[a]	C_1	C_2	C_{joint}	Joint	Rep	Joint	Rep	Joint	Rep
1.0	0	1.00	—	—	5.23	0.26	—	0.51	—	0.75	—
0.50	0.10	0.55	1.64	4.65	5.23	0.26	0.08	0.51	0.17	0.75	0.31
	0.05	0.53	1.96	4.50	5.23	0.26	0.09	0.51	0.21	0.75	0.36
	0.01	0.51	2.58	4.15	5.23	0.26	0.14	0.50	0.29	0.74	0.48
0.40	0.10	0.46	1.64	4.65	5.23	0.26	0.12	0.51	0.27	0.75	0.46
	0.05	0.43	1.96	4.50	5.23	0.26	0.14	0.50	0.30	0.74	0.51
	0.01	0.41	2.58	4.15	5.20	0.24	0.17	0.48	0.36	0.71	0.58
0.30	0.10	0.37	1.64	4.65	5.22	0.25	0.17	0.50	0.36	0.73	0.58
	0.05	0.34	1.96	4.50	5.21	0.24	0.18	0.48	0.37	0.72	0.60
	0.01	0.31	2.58	4.15	5.16	0.21	0.18	0.42	0.37	0.64	0.58
0.20	0.10	0.28	1.64	4.65	5.19	0.23	0.19	0.46	0.39	0.68	0.62
	0.05	0.24	1.96	4.50	5.16	0.21	0.18	0.42	0.38	0.63	0.59
	0.01	0.21	2.58	4.15	5.06	0.15	0.14	0.31	0.29	0.47	0.46

Shown is analysis of 1,000 cases and 1,000 controls, M = 300,000 markers, genome-wide significance level $\alpha_{genome} = 0.05$, multiplicative model, control risk allele frequency = 0.40 and prevalence = 0.10.
[a](Number of genotypes required for two-stage design) / (number of genotypes required when all markers are genotyped on all samples).