Biostatistics (2008), **9**, 2, *pp.* 201–215
doi:10.1093/biostatistics/kxm032
Advance Access publication on September 14, 2007

Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies

MITCHELL H. GAIL*

Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Boulevard, EPS 8032, Bethesda, MD 20892-7244, USA
gailm@mail.nih.gov

RUTH M. PFEIFFER

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD,USA

WILLIAM WHEELER, DAVID PEE

Information Management Services, Rockville, MD, USA

SUMMARY

Some case–control genome-wide association studies (CCGWASs) select promising single nucleotide polymorphisms (SNPs) by ranking corresponding p-values, rather than by applying the same p-value threshold to each SNP. For such a study, we define the detection probability (DP) for a specific disease-associated SNP as the probability that the SNP will be "T-selected," namely have one of the top T largest chi-square values (or smallest p-values) for trend tests of association. The corresponding proportion positive (PP) is the fraction of selected SNPs that are true disease-associated SNPs. We study DP and PP analytically and via simulations, both for fixed and for random effects models of genetic risk, that allow for heterogeneity in genetic risk. DP increases with genetic effect size and case–control sample size and decreases with the number of nondisease-associated SNPs, mainly through the ratio of T to N, the total number of SNPs. We show that DP increases very slowly with T, and the increment in DP per unit increase in T declines rapidly with T. DP is also diminished if the number of true disease SNPs exceeds T. For a genetic odds ratio per minor disease allele of 1.2 or less, even a CCGWAS with 1000 cases and 1000 controls requires T to be impractically large to achieve an acceptable DP, leading to PP values so low as to make the study futile and misleading. We further calculate the sample size of the initial CCGWAS that is required to minimize the total cost of a research program that also includes follow-up studies to examine the T-selected SNPs. A large initial CCGWAS is desirable if genetic effects are small or if the cost of a follow-up study is large.

*To whom correspondence should be addressed.

Keywords: Case–control study; Detection probability; Genetic association; Genome-wide association study; Ranking and selection; Whole genome scan.

1. INTRODUCTION

Case–control genome-wide association studies (CCGWASs) are used to detect associations of disease with genetic markers (single nucleotide polymorphisms [SNPs]) across the genome by comparing individuals with disease (cases) to disease-free individuals (controls). Several factors can lead to false-positive associations in CCGWASs, including population stratification and measurement error (Clayton *and others*, 2005). However, an overriding concern is the play of chance, when only a few true disease–associated SNPs are sought amidst the multitude of nondisease-associated SNPs.

One approach to control for multiplicity is to set stringent criteria for declaring an association statistically significant. For example, one might use the Bonferroni inequality to control the experiment-wise type I error rate. Two-stage designs have been proposed to reduce the amount of genotyping required while protecting the experiment-wise type I error rate (Skol *and others*, 2006). Less stringent but objective criteria such as the false discovery rate (Benjamini and Hochberg, 1995) have also been advocated.

Some investigators use p-values in a CCGWAS to rank and select promising SNPs for future study and are not concerned about the frequentist error control properties of the selection procedure. For example, the Cancer Genetic Markers of Susceptibility (CGEMS) project was designed to detect SNPs associated with prostate cancer (http://cgems.cancer.gov/). About 550 000 tagging SNPs were analyzed in an initial set of 1172 cases and 1157 controls (Yeager *and others*, 2007). About 28 000 most promising SNPs will be studied further in a second stage. These SNPs will be subjected to further winnowing in subsequent stages, leaving only 25–50 SNPs that will be regarded as sufficiently promising to warrant independent laboratory and epidemiologic investigations to attempt to establish a causal connection to disease. Altogether, data from 8400 cases and 8400 controls will be used.

The purpose of this paper is to investigate the probability that a specific disease SNP will be selected for further study in a CCGWAS and the probability that a selected SNP will be a true disease SNP. To be precise, we say that a SNP is "T-selected" or simply "selected" if its associated chi-square test statistic (or p-value) is among the top T chi-square test statistic values (or T lowest p-values). We call the probability that the test statistic for a specific disease SNP will be among the top T chi-square values in the sample the detection probability (DP). We also calculate PP, the proportion positive, namely the fraction of selected SNPs that are true disease–associated SNPs. The chi-square test we consider is a 2-sided trend test with additive (codominant) modeling of the SNP genotype (Armitage, 1955; Sasieni, 1997). Such a test does not require knowing whether the minor or major allele is associated with disease (Devlin and Roeder, 1999; Pfeiffer and Gail, 2003). We examine DP and PP both for the usual trend test based on the scores of the log-likelihood and for a Wald test.

It is computationally prohibitive to generate case–control data and perform 500 000 separate logistic analyses repeatedly. For the 1-stage design in which all SNPs are genotyped for every case and control, we develop asymptotic theory that allows us to study DP and PP in simulations, both for the score test and for the Wald test, and we also show how to calculate DP and PP analytically.

Our data give practical guidance as to required sample sizes and numbers of top ranks T needed to yield a high DP. We also study the effects of these parameters on PP, and on other factors crucial to designing a CCGWAS, including the rapid decrease in incremental DP per unit increase in T as T increases. Our findings give insight into how resources should be allocated between the CCGWAS and subsequent studies needed to follow up on the T-selected SNPs. In Section 4, we relate our work to the literature.

2. Materials and methods

2.1 *Study design and data for simulations*

We consider a population-based case–control design with n cases and n controls selected, respectively, at random from all cases and all controls in the source population. We assume that risk of disease is influenced by M out of N SNPs under study. For simulations, at each SNP $i = 1, 2, \ldots N$, we randomly and independently select a minor allele frequency, η_i, from the 299 686 minor allele frequencies that are 0.05 or greater in CGEMS (https://caintegrator.nci.nih.gov/cgems/downloadSetup.do). This minor allele frequency distribution had mean 0.2763, standard deviation 0.12, minimum 0.05, maximum 0.50, and quartiles 0.15 (25%), 0.26 (median), and 0.38 (75%). In each replicate of the simulations described below, minor allele frequencies were reassigned to each SNP in this way. CGEMS SNPs were chosen to be "tagging SNPs." Therefore, in the simulations in this paper, we regard the genotypes at the N SNPs as statistically independent in the source population, even though there may be correlation among nearby tagging SNPs.

2.2 *Logistic models*

Let $X_i = 0$, 1, or 2 be the number of minor alleles at locus i, let $\mathbf{X} = (X_1, X_2, \ldots, X_N)'$, and let $Y = 1$ or 0 for diseased or nondiseased subjects. Suppose SNPs $1, \ldots, M$ are associated with disease, while SNPs $M + 1, \ldots, N$ are not. Suppose in the source population, the probability of disease is given by $\text{logit}\{P(Y = 1|\mathbf{X})\} = \mu + \sum_{i=1}^{M} \beta_i X_i$, where $\text{logit}(u) = \log\{(u)/(1-u)\}$. For rare diseases (or for more common diseases over a confined age range such as 10 years), $P(Y = 1|\mathbf{X}) \doteq \exp(\mu + \sum_{i=1}^{M} \beta_i X_i)$. Assuming X_1 is independent of $X_2, X_3, \ldots, X_N$, we find $P(Y = 1|X_1) \doteq \exp(\mu^* + \beta_1 X_1)$, where $\mu^* = \mu + \log\{E \exp(\sum_{i=2}^{M} \beta_i X_i)\}$ and E is the expectation operator. For the case–control population, it follows that

$$\text{logit}\{P(Y = 1|X_1)\} \doteq \mu^{**} + \beta_1 X_1, \tag{2.1}$$

where $\mu^{**} = \mu^* + \log(\pi_1/\pi_0)$, π_1 is the proportion of cases in the source population that are in the case–control study, and π_0 is the analogous proportion for controls.

2.3 *Models for the genetic effect β*

For nondisease-associated SNPs, $\beta_i = 0$. For disease-associated SNPs, we consider a random effects model and a fixed effects model for β_i, $i = 1, \ldots, M$. Under the random effects model, each $\beta_i, i = 1, \ldots, M$, is drawn independently from a normal distribution with mean 0 and variance τ^2. As $E|\beta_i| = \tau(2/\pi)^{1/2} \doteq 0.798\tau$, large values of τ^2 correspond to large average effects of disease SNPs. We also consider fixed effects models $\beta_i = \beta$, for $i = 1, \ldots, M$, for a fixed β.

2.4 *Properties of the parameter estimates and chi-square tests*

Maximum likelihood estimation for a cohort study applied to case–control data with (2.1) yields a fully efficient estimate of β_1 and a consistent variance estimate, $\hat{\text{V}}\text{ar}(\hat{\beta}_1)$ (Prentice and Pyke, 1979). We show that if genotypes are independent in the source population, as we assume, and if the disease is rare, then genotypes are independent in the samples of cases and of controls. For simplicity, we prove independence for 2 genotypes, but the result extends to N genotypes. Consider a fixed set of parameters $\beta = (\beta_1, \beta_2, \ldots, \beta_N)'$. Let $\rho_{ki} = P(X_i = k)$ be the probability that the genotype at locus i has k minor alleles in the source population. Under independence of genotypes and the rare disease assumption, $P(Y = 1|X_i; \mu_i^*, \beta_i) \doteq \exp(\mu_i^* + \beta_i X_i)$ in the source population. Thus,

$$f_{ki} \equiv P(X_i = k|Y = 1) \doteq \rho_{ki} \exp(\beta_i k) \left\{ \sum_{l=0}^{2} \rho_{li} \exp(\beta_i l) \right\}^{-1} \quad \text{and}$$

$$g_{ki} \equiv P(X_i = k|Y = 0) \doteq \rho_{ki}$$

in the case–control sample. Likewise,

$$\begin{aligned} f_{klih} &\equiv P(X_i = k, X_h = l|Y = 1) \\ &\doteq \rho_{ki}\rho_{lh} \exp(\beta_i k + \beta_h l) \left\{ \sum_{s_1=0}^{2} \sum_{s_2=0}^{2} \rho_{s_1 i}\rho_{s_2 h} \exp(\beta_i s_1 + \beta_h s_2) \right\}^{-1} \\ &= f_{ki} f_{lh} \end{aligned}$$

and

$$g_{klih} \equiv P(X_i = k, X_h = l|Y = 0) \doteq \rho_{ki}\rho_{lh}.$$

Thus, X_i and X_j are independent, not only in controls but also in cases.

It follows that each SNP can be analyzed separately based on (2.1), resulting in independent chi-square statistics across SNPs. One chi-square test for $\beta_i = 0$ is the Wald statistic $C_i = \hat{\beta}_i^2 / \hat{\text{V}}\text{ar}(\hat{\beta}_i)$. A second chi-square test for the ith SNP is the score test (Armitage, 1955) $CS_i \equiv U_i^2 / \hat{\text{V}}\text{ar}_0(U_i)$, where $U_i = 0.5(\sum_{\text{cases},1}^{n} X_i - \sum_{\text{controls},1}^{n} X_i)$, and $\text{Var}_0(U_i)$ is the null variance of U_i (supplementary Appendix available at *Biostatistics* online [http://www.biostatistics.oxfordjournals.org]) . In the expression for U_i, the index i in X_i refers to the locus and not to the cases or controls.

2.5 *Simulations and ranking criteria*

We use the asymptotic normal distributions of $\hat{\beta}_i$ and U_i (see supplementary Appendix available at *Biostatistics* online [http://www.biostatistics.oxfordjournals.org]) to generate realizations of these quantities rapidly in GAUSS (Aptec Systems, 2005). For givenβ_i, we calculate $\text{Var}(\hat{\beta}_1)$ by taking the expectation of the prospective information matrix, I, from (2.1), with respect to the retrospective sampling distribution. Letting

$$p_x \equiv 1 - q_x = P(Y = 1|X_i = x; \mu_i^{**}, \beta_i), \text{ we find } I_{11} = E(p_x q_x) = n \sum_{x=0}^{2} (f_{xi} + g_{xi}) p_x q_x,$$

$$I_{21} = I_{12} = E(X p_x q_x) = n \sum_{x=0}^{2} (f_{xi} + g_{xi}) x p_x q_x, \text{ and } I_{22} = E(X^2 p_x q_x) = n \sum_{x=0}^{2} (f_{xi} + g_{xi}) x^2 p_x q_x.$$

We obtain μ_i^{**}in these equations from $0.5 = 0.5 \sum_{x=0}^{2} (f_{xi} + g_{xi}) P(Y = 1|X_i = x; \mu_i^{**}, \beta_i)$.

Conditional on $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_N)'$, $\text{Var}(\hat{\beta}_i) \equiv \sigma_i^2(\beta_i) = I_{22.1}^{-1} \equiv (I_{22} - I_{21} I_{11}^{-1} I_{12})^{-1}$. For nondisease loci and for disease loci under the fixed effects disease model, the conditional variances equal the unconditional variances. Although the results in Sections 2.4 and 2.5 hold for generalρ_{ki}, we assume Hardy–Weinberg Equilibrium in simulations. Thus, $\rho_{0i} = (1 - \eta_i)^2$, $\rho_{1i} = 2\eta_i(1 - \eta_i)$, and $\rho_{2i} = \eta_i^2$.

For each of the various parameter settings, we generate NSIM = 10 000 independent simulations. From the previous theory, we can generate C_i as follows. For the fixed effects model, set $\beta_i = \beta$ for $i = 1, 2, \ldots, M$ and $\beta_i = 0$ for $i = M + 1, \ldots, N$. For the random effects model, draw β_i from the normal distribution $N(0, \tau^2)$ for $i = 1, 2, \ldots, M$. For $i = M+1, \ldots, N$, set $\beta_i = 0$. Under either model, draw an independent random allele frequency η_i for each SNP, and conditional on the values of β_i and η_i, compute $\sigma_i^2(\beta_i)$. Then draw $\hat{\beta}_i$ from $N(\beta_i, \sigma_i^2(\beta_i))$ and compute $C_i = \hat{\beta}_i^2 / \sigma_i^2(\beta_i)$. This quantity has

the same asymptotic distribution as $C_i = \hat{\beta}_i^2/\hat{\mathrm{V}}\mathrm{ar}(\hat{\beta}_i)$. A similar approach is used to generate a quantity that is asymptotically equivalent to $CS_i \equiv U_i^2/\hat{\mathrm{V}}\mathrm{ar}_0(U_i)$, as described in the supplementary Appendix available at *Biostatistics* online (http://www.biostatistics.oxfordjournals.org).

Define $I(m, \mathrm{ISIM}, T) = 1$ if rank $(C_m) > N - T$, namely SNP m has a test statistic C_m that is in the top T ranks of the N-ranked values of C_i in simulation ISIM, and 0 otherwise. Here, m indexes one of the M disease SNPs. The DP is estimated by

$$\hat{\mathrm{DP}} = \mathrm{NSIM}^{-1} M^{-1} \sum_{m=1}^{M} \sum_{\mathrm{ISIM}=1}^{\mathrm{NSIM}} I(m, \mathrm{ISIM}, T). \tag{2.2}$$

The inner summation divided by NSIM is the proportion of simulations in which $C_m$is in the top T ranks. Because the disease SNPs are exchangeable under either the random effects or the fixed effects model, the average of this quantity over m yields the average probability that a disease SNP will be found in the top T ranks. Thus, $\hat{\mathrm{DP}}$ is an estimate of the probability that a given disease SNP will have an associated Wald test in the top T ranks. By exchanging the order of summation in (2.2), we see that $\hat{\mathrm{DP}}$ has an alternative interpretation as the average proportion of disease SNPs that are in the top T ranks. Therefore, PP can be estimated from $\hat{\mathrm{PP}} = (M)(\hat{\mathrm{DP}})/T$. We evaluate the rankings of the score test by replacing C_m with CS_m and C_i with CS_i in the definition of $I(m, \mathrm{ISIM}, T)$. We study $T = 25, 50, 100, 250, 500, 1000, 5000, 10\,000$, and $25\,000$, which, when divided by the total number of SNPs, $N = 500\,000$, yields respective fractions 0.00005, 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.01, 0.02, and 0.05.

2.6 *Analytic calculation of DP*

For all nondisease-associated SNPs, the asymptotic distribution of the Wald test and of the score test is a central chi-square distribution with 1 degree of freedom, F, even if their allele frequencies vary. For disease-associated SNPs, however, the asymptotic distributions of these test statistics vary under the random effects model, or if their allele frequencies vary. For fixed η_i and β_i, let G_i be the distribution of C_i for $i = 1, 2, \ldots, M$. Consider a particular disease SNP, without loss of generality, SNP 1. Let $g_1(c)$ be the density of C_1, and let $H_1(c)$ be the event that C_1 is in the interval $[c, c + dc)$. Define $H_2(m; c, M)$ to be the event that m of the remaining $M - 1$ disease SNPs have C_i values greater than c, and let $H_3(T - m - 1; c, m)$ be the event that no more than $T - 1 - m$ nondisease SNPs have C_i values greater than c. Note that the intersection of these 3 events implies that C_1 is in the top T ranks. Thus, conditionally on η_i and β_i, DP is given by

$$\mathrm{DP} = \int_0^\infty \left[\sum_{m=0}^{\min(M-1,T-1)} P(H_2(m; c)|c) \sum_{s=0}^{T-1-m} \binom{N-M}{s} \{1 - F(c)\}^s \{F(c)\}^{N-M-s} \right] g_1(c) dc, \tag{2.3}$$

where F is the central chi-square distribution with 1 degree of freedom. Equation (2.3) is the integral over c of $P\{H_1(c)\}P\{H_2(m; c, M)|H_1(c)\}P\{H_3(T - m - 1; c, m)|H_1(c), H_2(m; c, M)\}$. If $M = 1$, $P\{H_2(m = 0; c, M)\} = G_1(c)$ and $P\{H_2(m = 1; c, M)\} = 1 - G_1(c)$. If $M > 1$, the following recursion, similar to that in Gail *and others* (1979), can be used to calculate the quantity $P(H_2(m; c)|c)$ in (2.3): $P\{H_2(m; c, M)\} = G_M(c)P\{H_2(m; c, M-1)\} + \{1 - G_M(c)\}P\{H_2(m-1; c, M-1)\}$. Initial conditions are $P\{H_2(m = 0; c, M = 0)\} = 1$, $P\{H_2(m = 1; c, M = 0)\} = 0$, and $P\{H_2(m = -1; c, M\} = 0$ for all M. The unconditional value of DP is obtained by averaging (2.3) over the distribution of η_i and, under the random effects model, over β_i, for $i = 1, 2, \ldots, M$.

If all the disease loci have the same distribution, $G(c)$, (2.3) simplifies to

$$\mathrm{DP} = \int_0^\infty \left[\sum_{m=0}^{\min(M-1,T-1)} \binom{M-1}{m} g(c) G(c)^{M-1-m} \{1-G(c)\}^m \times \sum_{s=0}^{T-1-m} \binom{N-M}{s} \{1-F(c)\}^s \{F(c)\}^{N-M-s} \right] dc. \tag{2.4}$$

Expressions (2.4), (2.5), and (2.6) apply equally to CS_i. We used (2.4) with a fixed allele frequency and common fixed effect or a common random effect for all disease loci to check the simulation procedures and to compute DP and PP in these special cases. For fixed η and β, C_i has distribution $G(c) = G^*(c; \beta^2/\sigma_i^2(\beta))$, where G^* is a noncentral chi-square distribution with 1 degree of freedom and noncentrality $\beta^2/\sigma_i^2(\beta)$. For the random effects model, G is the average of G^*over the distribution of β. For fixed η and β, CS_i has distribution $G(c) = G^*(c\{\mathrm{Var}_{\beta_i=0}(U_i)/\mathrm{Var}_{\beta_i}(U_i)\}; \{\delta(\beta)\}^2/\mathrm{Var}_\beta(U_i))$, where $\delta(\beta)$ and $\mathrm{Var}_\beta(U_i)$ are given in the supplementary Appendix available at *Biostatistics* online (http://www.biostatistics.oxfordjournals.org). For the random effects model, G is the average of G^*over the distribution of β.

For the fixed effects model with fixed allele frequencies $\eta_i = \eta$, G for the Wald test has noncentrality parameter $\mathrm{NCP} = \beta^2/\sigma^2(\beta)$. If η varies in a fixed effects model or if β arises from a random effects model, there is no parameter, NCP, that characterizes G for the Wald test. For descriptive purposes for such cases, we use the "approximate NCP" defined by $\beta^2 E\{1/\sigma^2(\beta)\}$, in which the expectation is estimated analytically for fixed $\eta_i = \eta$ and empirically in simulations otherwise.

An excellent approximation to (2.4) for $T > 20$ that requires no integration is

$$\frac{\sum_{m=0}^{\min(T,M)} (m/M) \binom{M}{m} G(q_m)^{M-m} \{1-G(q_m)\}^m}{\sum_{m=0}^{\min(T,M)} \binom{M}{m} G(q_m)^{M-m} \{1-G(q_m)\}^m}, \tag{2.5}$$

where $q_m = F^{-1}\{(N-M-T+m)/(N-M+1)\}$. For $M = 1$, this approximation is very nearly

$$1 - G\{F^{-1}(1-T/N)\}. \tag{2.6}$$

2.7 *Analytic calculation of PP*

The PP is the expected number of true disease SNPs among the T-selected SNPs, divided by T, namely

$$\mathrm{PP} = (M/T)(\mathrm{DP}). \tag{2.7}$$

Thus, PP can be calculated analytically from the previous calculations for DP.

3. RESULTS

3.1 *Detection probability*

Figure 1 shows estimates of DP for the fixed effects model from 10 000 simulations for the score test with $M = 1$ and $N = 500\,000$ for a case–control study with $n = 1000$ cases and $n = 1000$ controls;

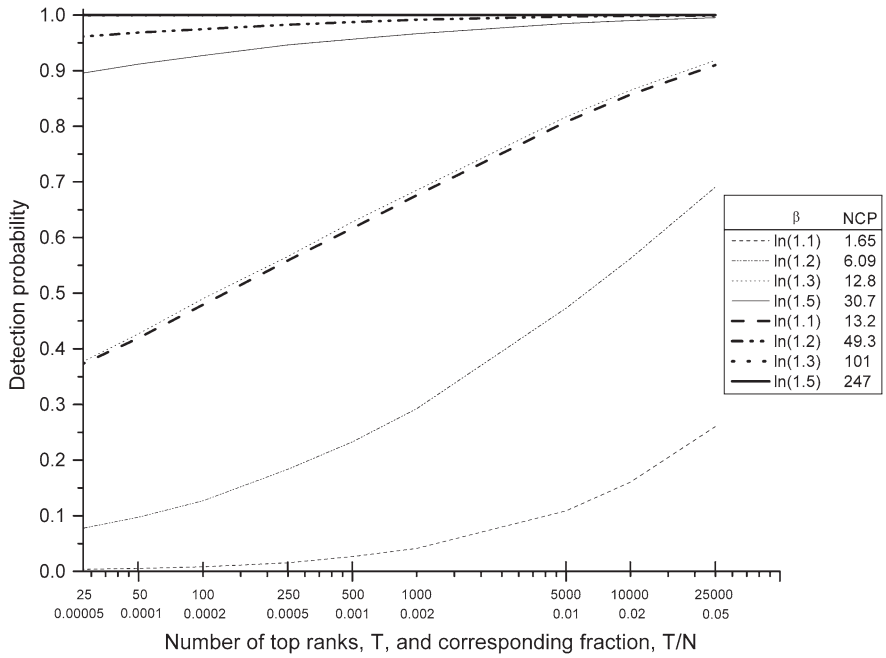

Fig. 1. Estimated DP for the score test and for the fixed effects model plotted against $\log(T)$ based on 10 000 simulations with minor allele frequencies drawn at random from the CGEMs distribution. Other parameters are $N = 500\,000$, $M = 1$ disease allele, and $n = 1000$ cases and controls (not bold) or $n = 8000$ cases and controls (bold). Approximate NCPs are computed as β^2 times the average value of $\{\mathrm{Var}(\hat{\beta})\}^{-1}$ in the simulations.

results are also shown for $n = 8000$ cases and controls. For $n = 1000$, DP is near 1.0 for $\beta = \log(2.0)$ even for $T = 25$ (not shown). For $\beta = \log(1.2)$, DP is only 0.07 at $T = 25$ and rises gradually to 0.69 at $T = 25\,000$. Because T is plotted on a log scale, it is apparent that DP increases very slowly with increasing T. For $\beta = \log(1.2)$ and $n = 8000$ cases and controls, DP = 0.96 at $T = 25$ and DP = 0.999 at $T = 25\,000$. Approximate NCP values are shown for the 8 loci in Figure 1.

Figure 2 shows estimates of DP for the random effects model. Even though values of τ were chosen such that the mean absolute value of β in the random effects model equaled the value of β in the corresponding fixed effects model, the DP of the score test is less in the random effects model for large τ (and large NCP). For example, with $n = 1000$ and $\tau = \log(2.0)/0.798$, the DP increases from 0.71 at $T = 25$ to 0.85 at $T = 25\,000$, whereas, in the fixed effects model with $\beta = \log(2.0)$ the DP is 1.000 for $T = 25$ (locus not shown). For small NCP, the DP of the random effects model exceeds that of a fixed effects model with comparable NCP.

Figures 1 and 2 can be used to assess DP for other β values by interpolation. One can also interpolate based on NCP values; in particular, for other sample sizes n^* and genetic effects β^*, estimate DP from some choice of n and β in Figure 1 satisfying $n^*(\beta^*)^2 = n\beta^2$. The approximations (2.5) and (2.6) indicate that DP depends on T and N mainly through T/N. Unreported numerical examples show that this is also true for simulated estimates, such as those in Figures 1 and 2. Thus, Figures 1 and 2 can be used for other values, say N^* and T^*, by referring to the value of T that satisfies $T/N = T^*/N^*$. Results for the Wald test are visually indistinguishable from those in Figures 1 and 2 for the score test and are not shown.

The number of competing disease loci M has little effect on DP provided $M < T$, but for $M > T$, DP declines sharply with increasing M. This phenomenon was demonstrated in simulations that allow for

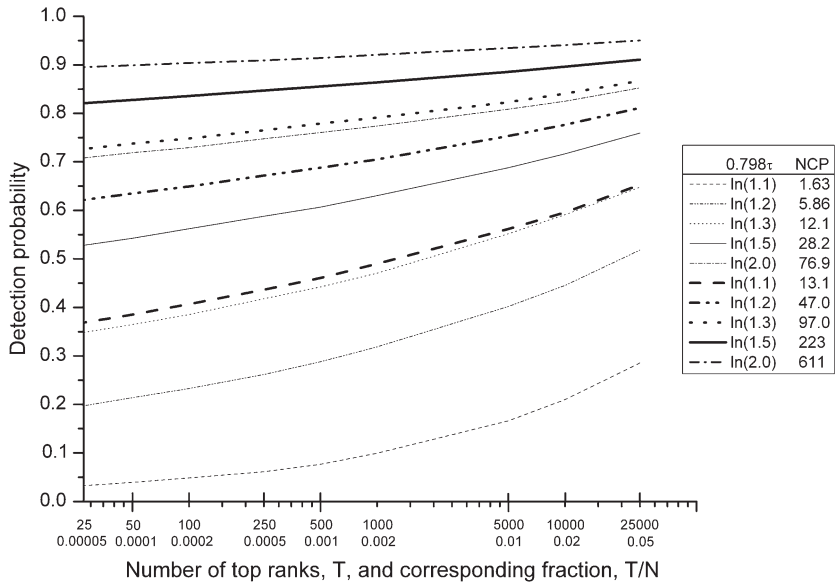

Fig. 2. Estimated DP for the score test and for the random effects model plotted against $\log(T)$ based on 10 000 simulations with minor allele frequencies drawn at random from the CGEMs distribution. Other parameters are $N = 500\,000$, $M = 1$ disease allele, and $n = 1000$ cases and controls (not bold) or $n = 8000$ cases and controls (bold). Approximate NCPs are computed as β^2 times the average value of $\{\mathrm{Var}(\hat{\beta})\}^{-1}$ in the simulations, where $\beta = \log(1.1)$, $\log(1.2)$, $\log(1.3)$, $\log(1.5)$, or $\log(2)$ correspond to the values of the standard deviation of the random effects distribution, $\tau = \log(\beta)/0.798$.

variable allele frequencies (data not shown) as well as by analytic results (scenarios 4–6 in Table 1) for the random effects model with $\tau = \beta/0.798 = \log(1.2)/0.798$, $n = 8000$ and η fixed at the mean of the CGEMS distribution, 0.2673. DP for $M = 100$ is much reduced compared to $M = 1$ or $M = 10$ when $T < 100$, and DP is much less for $M = 10$ than for $M = 1$ with $T = 1$. Similar results hold for the fixed effects model (supplementary Table S1 available at *Biostatistics* online [http://www.biostatistics.oxfordjournals.org]).

For SNPs with fixed minor allele frequencies $\eta_i = \eta$, DP is lower for $\eta = 0.05$ than for $\eta = 0.50$, as seen from scenarios 8 and 9 in Table 1 for the random effects model and supplementry Table S1 available at *Biostatistics* online (http://www.biostatistics.oxfordjournals.org) for the fixed effects model. From supplementary Appendix equation (A.8) available at *Biostatistics* online (http://www.biostatistics.oxfordjournals.org), the NCP for $\eta = 0.5$ is greater than that for $\eta = 0.05$ by a factor $(0.5)^2/(0.05)(0.95) = 5.26$.

Very large values of T may be needed to ensure a DP above 0.5 with $n = 1000$ (Figures 1 and 2). However, it may not be feasible to evaluate many loci in independent data or to develop functional assays or knockout systems to confirm their importance. It is therefore important to study the increment in DP per unit increase in T in order to determine how rapidly the returns from increasing T diminish. Figure 3, based on (2.4), plots the increment in DP for the random effects model from $T - 1$ to T against $\log(T)$, for $T = 2, 3, \ldots, 500$ and for various choices of $\tau = \beta/0.798$ and sample sizes $n = 1000$ or $n = 8000$ cases and controls. For $M = 1$ (not bold loci), with $n = 8000$ and high NCP, there is little to be gained by increasing T beyond 100; for $M = 1$ and $n = 1000$ with smaller NCP, little is gained by increasing T beyond 500. For $M = 100$ (bold loci) and $n = 8000$ with large NCP, increases in T are

Table 1. *DP and PP for the Wald test and for random effects model*[†]

Case	0.798τ	M	η	n	Method[‡]	$T = 1$		$T = 10$		$T = 100$		$T = 1000$	
						DP	PP	DP	PP	DP	PP	DP	PP
1	log(1.2)	1	0.2673	1000	A	0.143	0.143	0.198	0.020	0.263	0.003	0.353	0.000
2	log(1.2)	10	0.2673	1000	A	0.078	0.781	0.193	0.193	0.262	0.026	0.353	0.004
3	log(1.2)	100	0.2673	1000	A	0.010	1.000	0.099	0.986	0.254	0.254	0.352	0.035
4	log(1.2)	1	0.2673	8000	A	0.592	0.592	0.638	0.064	0.682	0.007	0.734	0.001
5	log(1.2)	10	0.2673	8000	A	0.100	1.000	0.622	0.622	0.681	0.068	0.734	0.007
6	log(1.2)	100	0.2673	8000	A	0.010	1.000	0.100	1.000	0.661	0.661	0.732	0.073
7	log(2.0)	1	0.2673	1000	A	0.685	0.685	0.722	0.072	0.758	0.008	0.798	0.001
8	log(1.2)	1	0.05	1000	A	0.010	0.010	0.022	0.002	0.045	0.000	0.096	0.000
9	log(1.2)	1	0.50	1000	A	0.191	0.191	0.250	0.025	0.318	0.003	0.407	0.000
10	log(1.2)	1	Random	1000	S	0.126	0.126	0.172	0.017	0.231	0.002	0.318	0.0003
11	log(1.2)	100	Random	1000	S	0.010	1.000	0.097	0.971	0.227	0.227	0.320	0.032

[†]Calculations based on $N = 500\,000$ loci. The allele frequency η is fixed as shown except for cases 10 and 11, in which allele frequencies were drawn from the CGEMS minor allele frequency distribution, with mean 0.2673. Each of the M disease loci has a trend effect β drawn independently from a normal distribution with mean 0 and variance τ^2.

[‡]A indicates that the calculations were based on analytical equations (2.4) and (2.7) under the assumption that the Wald tests for disease SNPs have common distribution given by supplementary Appendix equation (A.10) available at *Biostatistics* online (http://www.biostatistics.oxfordjournals.org). S indicates that the estimates were from simulations with 10 000 repetitions that allow variation in allele frequencies and disease SNP log relative odds.

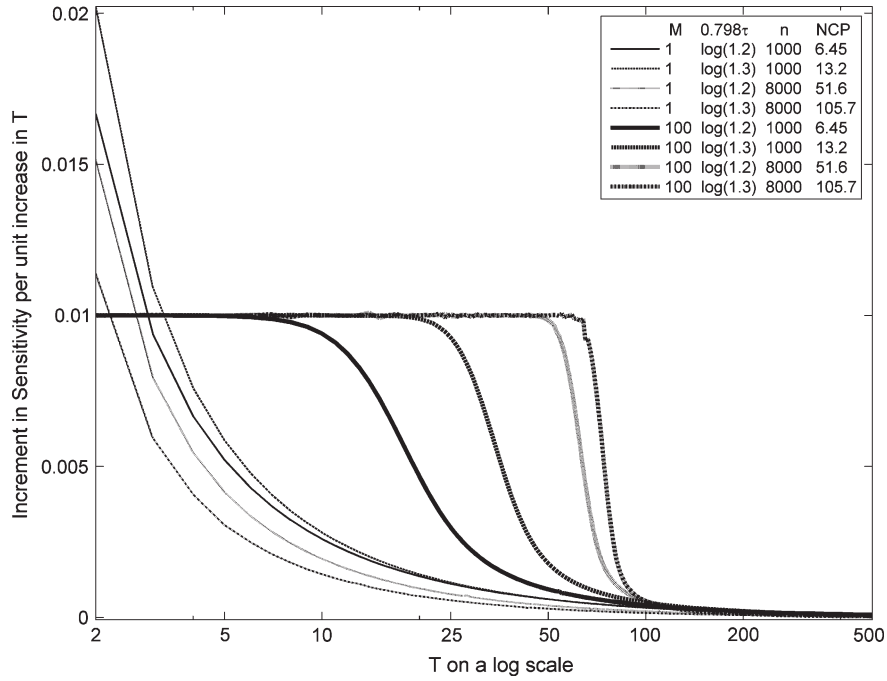

Fig. 3. Increment in DP per unit increase in T for the Wald test and for the random effects model, plotted against $\log(T)$ for various values of the standard deviation of the random effects distribution, $\tau = \log(\beta)/0.798$, for $n = 1000$ or 8000 cases and controls, and for $M = 1$ disease locus (not bold) and $M = 100$ disease loci (bold). Approximate NCPs are shown and were computed as β^2 times the expectation over β of $\{\mathrm{Var}(\hat{\beta})\}^{-1}$. We assume $N = 500\,000$ and calculate the DP from (2.4) for minor allele frequency fixed at $\eta = 0.2673$, the mean of the CGEMS distribution.

worthwhile for $T \leqslant 100$, but not beyond $T = 200$; for $M = 100$ and $n = 1000$ with smaller NCP, values of T in the range 200–500 still yield some incremental increases in DP. Increments in DP decrease even more rapidly for the fixed effects model (supplementry Figure S1 available at *Biostatistics* online [http://www.biostatistics.oxfordjournals.org]).

3.2 *Proportion positive*

Factors influencing PP are depicted in Figure 4 for the random effects model, based on (2.4) and (2.7). PP is higher for $M = 100$ than for $M = 1$, for every choice of β , n, and T. For $T < M$, a PP near 1.0 is achieved for DP near 1.0, namely with large NCP. As anticipated from (2.7), PP decreases as T increases for $T > M$ (Figure 4). For $M = 100$ (bold loci) and $T < M$, PP is near 1.0 for $n = 8000$ (large NCP and hence large DP), but falls to below $M/T = 0.004$ at $T = 25\,000$. Even for $n = 1000$, for $M = 100$ and $T = 10$, PP is near 1.0. Thus, if there are $M = 100$ disease loci, the chance that a randomly selected SNP in the top 10 selected SNPs is a disease SNP is nearly 1.0. For $M = 1$ (not bold loci), PP falls rapidly for $T > 1$. If the NCP (and therefore the DP) is large, the PP is near 0.7 for $T = 1$, but falls to below $1/T$ as T increases. Figure 4 and Table 1 illustrate that raising T to increase DP decreases PP for $T > M$. The PP curves for a fixed effects model (supplementary Figure S2 available at *Biostatistics* online [http://www.biostatistics.oxfordjournals.org]) have similar shapes as for the random effects model (Figure 4). For large comparable NCP values, the fixed effects model has higher PP, but for models with comparable low NCP values, the random effects model has higher PP (compare Figure 4 with supplementary

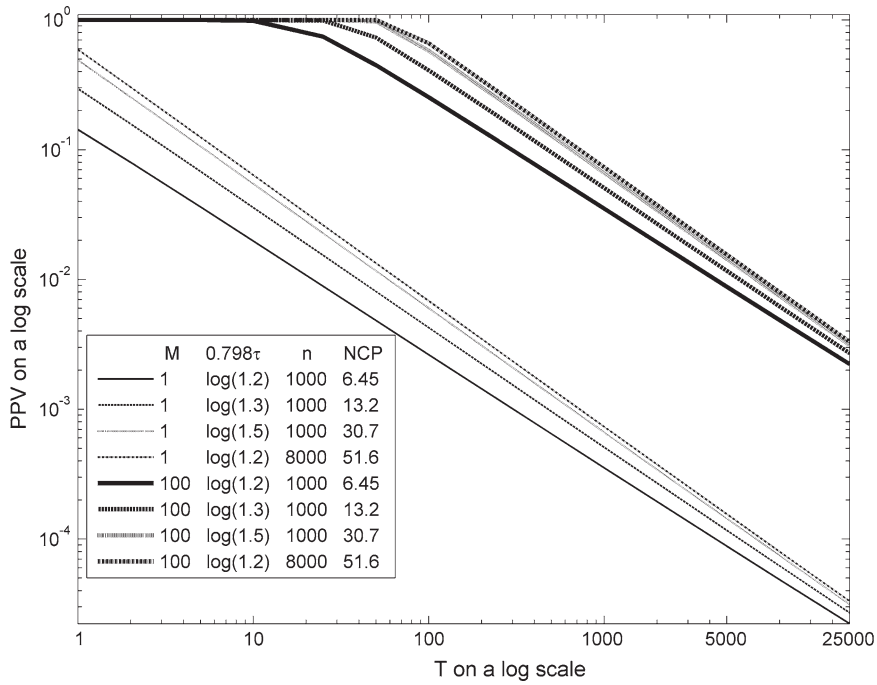

Fig. 4. PP on log scale for the Wald test and for the random effects model plotted against $\log(T)$ for $M = 1$(not bold) or $M = 100$ (bold) and for various values of the standard deviation of the random effects distribution, $\tau = \log(\beta)/0.798$, and numbers of cases and controls, n = 1000 or 8000. Approximate NCPs are shown and were computed as β^2 times the expectation over β of $\{\mathrm{Var}(\hat{\beta})\}^{-1}$. Other parameters are $N = 500\,000$ and the minor allele frequency, fixed at $\eta = 0.2673$, the mean of the CGEMS distribution.

Figure S2 available at *Biostatistics* online [http://www.biostatistics.oxfordjournals.org] and Table 1 with supplementary Table S1 available at *Biostatistics* online [http://www.biostatistics.oxfordjournals.org]).

3.3 *Designs to minimize total cost*

Suppose the total cost to identify and confirm the importance of a potential disease SNP is $2nC_1 + C_2T$, where C_1 is the cost to recruit and genotype a subject for a CCGWAS and C_2 is the cost of a confirmatory study on each of the T-selected candidate SNPs. We call C_1 the "initial study cost" and C_2 the "follow-up cost," which might require laboratory investigations of functionality or confirmatory association studies. As n increases, the initial study cost increases linearly, but the number T required to achieve a desired DP decreases, reducing the follow-up cost. Expressed in units of C_1, the total cost is $2n + (C_2/C_1)T$. We consider cost ratios, $C_2/C_1 = 10$ or 1000. For example, if the initial study cost is $C_1 = \$1000$ per subject, a laboratory study to check the functionality of a locus might cost $C_2 = \$10\,000$, and a confirmatory epidemiologic study might cost \$1 000 000, corresponding to cost ratios 10 and 1000. As an example, we consider fixed effects models with $\beta = \log(1.2)$ (bold loci) or $\beta = \log(1.3)$ (not bold loci), with $M = 1$ or $M = 100$, with minor allele frequency fixed at the mean CGEMS value, $\eta = 0.2673$, and with $N = 500\,000$. For each fixed n, we can invert (2.4) to find the T that gives DP $= 0.9$, and for various cost ratios find that value of n that minimizes total cost (Figure 5). Total costs are higher for $\beta = \log(1.2)$, because larger values of n and T are needed to yield DP = 0.9 (Figure 1). For $\beta = \log(1.3)$ and $M = 1$,

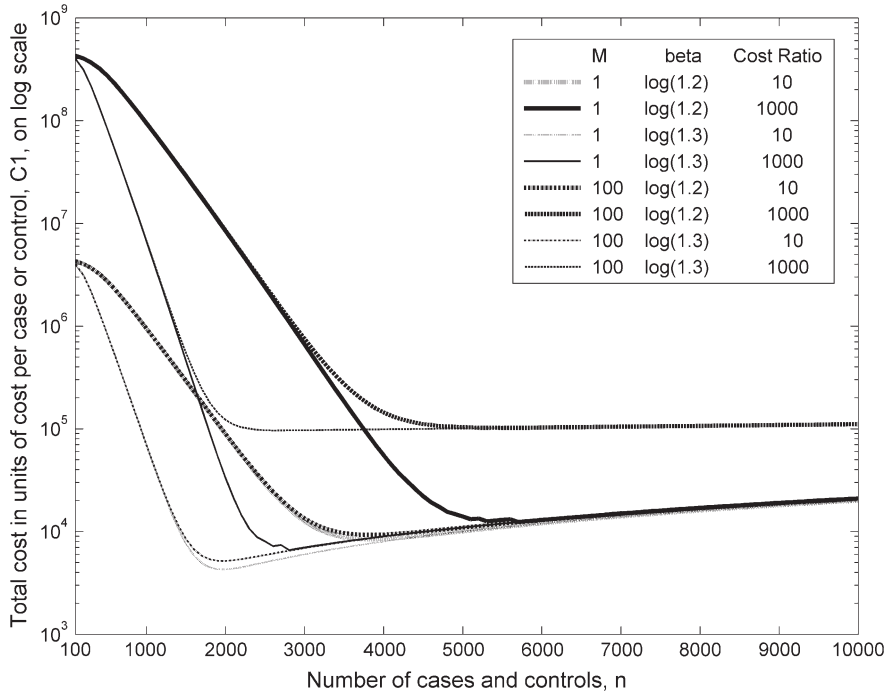

Fig. 5. Total cost, in units of C_1, the cost per subject in the CCGWAS, plotted as a function of n, the numbers of cases and controls for the Wald test and for a fixed effects model. Cost ratios $C_2/C_1 = 10$ and 1000 are studied, where C_2 is the cost of a follow-up study, and the fixed effects were $\beta = \log(1.2)$ (bold) or $\beta = \log(1.3)$ (not bold). Other parameters were $N = 500\,000$, $M = 1$ disease allele, and the minor allele frequency, fixed at $\eta = 0.2673$, the mean of the CGEMS distribution.

the values of (n, T) that minimized total cost were (1948, 38) and (2711, 1), respectively, for cost ratios 10 and 1000; the corresponding total costs were 4276 and 6422, in units of C_1. For $\beta = \log(1.3)$ and $M = 100$, the values of (n, T) that minimized total cost were (1944, 128) and (2598, 91) with respective total costs 5168 and 96 196. The optimal CCGWAS size is little affected by having $M = 100$ instead of $M = 1$, but the total cost can be much larger because larger numbers, T, are needed to assure DP = 0.9. For $\beta = \log(1.2)$ and $M = 1$, the values of (n, T) that minimized total cost were (3810, 77) and (5673, 1) with respective total costs 8390 and 12 346. If $M = 100$ instead, the results are (3801, 168) and (5436, 91) with respective total costs 9282 and 101 872. These calculations indicate that the optimal n increases with decreasing genetic effect size and with increasing cost ratio. The optimal T increases with decreasing genetic effect and with the number of disease genes, M, but decreases with increasing cost ratio.

4. DISCUSSION

We studied DP, the probability that a given disease locus will be T-selected based on the largest T chi-square values (or corresponding smallest p-values). DP has the alternative interpretation as the proportion of exchangeable disease loci that will be T-selected. We have shown how DP can be calculated based on the underlying logistic risk model in the source population and on the case–control sampling. We calculated DP analytically as the average of (2.3) over the distribution of allele frequencies, and, for random effects models that allow for heterogeneity of disease SNP effects, over the corresponding log odds ratios. To

study realistic designs, we simulated the results in Figures 1 and 2 based on the distribution of allele frequencies in CGEMS. We showed that if genotypes are independent in the source population and the disease is rare, chi-square tests for individual SNPs are independent, and we presented asymptotic theory that permits one to generate realizations for simulations rapidly.

A number of factors affect DP, especially the magnitude of the disease SNP effect, β (Figures 1 and 2). DP depends on the number of nondisease loci, $N - M$, but mainly through the ratio T/N; large N may require very large T to insure high DP. Competition among multiple disease loci ($M > 1$) can reduce DP, but this effect is only large when $M > T$. DP is less in SNPs with small minor allele frequencies.

Our assumption that the SNP genotypes are independent is valid if the original $N = 500\,000$ tagging SNPs are independent. Zaykin and Zhivotovsky (2005) showed that correlations of p-values within linkage disequilibrium blocks of SNPs or among such blocks have little effect on selection probabilities similar to DP, because such correlations do not extend beyond a small portion of the genome. It is likely, therefore, that DP is also little affected by such correlations.

Our work complements and differs from that of Zaykin and Zhivotovsky (2005). First, our criterion, DP, is not the same as the criteria they evaluated, namely the probability that all disease SNPs would have p-values lower than that of the ith smallest p-value among nondisease SNPs (their equation A.6) and the probability that at least one disease SNP will have a p-value below that of the ith smallest p-value among nondisease SNPs (their equation A.3). Although each of these criteria may be useful in some circumstances, we believe that DP gives the best overall assessment of the ability to detect disease SNPs. Second, we relate DP to the parameters in a logistic risk model with case–control sampling.

Satagopan *and others* (2004) studied ranking procedures in 1- and 2-stage designs and computed the probability that at least a desired number, m, of the M disease SNPs would have normally distributed trend tests exceeding the largest normally distributed trend test among nondisease SNPs. This criterion differs from DP except in the case $T = M = m = 1$. Moreover, the calculations assume that it is known whether the minor or major allele is positively associated with disease. Our results are based on chi-square statistics that are invariant to this polarity, and are therefore more suited to exploratory CCGWASs.

Wacholder *and others* (2004) made recommendations regarding alpha levels and power required to control the "false-positive report probability" (FPRP), namely the probability that a genetic variant selected on the basis of rejecting the null hypothesis of no association would be a false discovery. The FPRP concept, however, does not require any consideration of how many SNPs are examined. Each is considered on its own. In contrast, in the present paper we define "selection" not in terms of a hypothesis test but in terms of a ranking of chi-square tests (or p-values) for all SNPs. Because the rank selection criterion depends not only on the chi-square test for a given SNP but also on the chi-square tests for all other SNPs, we have used the terms DP and PP rather than the analogous terms, sensitivity or true positive fraction (Pepe, 2003) and positive predictive value (Vecchio, 1966), that are used for diagnostic tests in independent subjects.

Our methods give insight into aspects of CCGWAS design. Low PP was found for studies with modest effect sizes or case–control sample sizes, illustrating the futility of trying to identify disease loci from a CCGWAS with small n. Even for $n = 1000$ cases and controls, one requires $T = 4710$ to achieve DP $= 0.5$ for a fixed effect $\beta = \log(1.2)$ with $M = 1$ and fixed $\eta = 0.2673$. The corresponding PP $= 0.0001$. For $M = 100$, $T = 4758$ is required to attain DP $= 0.5$, and PP $= 0.0105$ in this case. Thus, the vast majority of selected SNPs will be false positives. If $n = 8000$ cases and controls are used instead and $M = 1$, then $T = 1$ yields DP $= 0.992$ and the corresponding PP $= 0.992$. For $M = 100$, $T = 50$ is required to yield DP $= 0.5$, and the corresponding PP $= 1.00$. Thus, large samples are needed to detect a disease SNP with odds ratio 1.2 reliably with a value of T that is practical for follow-up studies. CCGWASs with only a few hundred cases and controls will have low DP unless T is extravagantly large, in which case the PP will be very small, necessitating too many follow-up studies to be feasible.

Our methods also provide information on how to allocate resources in a research program that supports both discovery of promising SNPs in an initial CCGWAS and follow-up studies to evaluate initial leads among the T-selected SNPs. When follow-up costs for each selected SNP are 10 or 1000 times the cost of a subject in the CCGWAS, or when genetic effects are small, then the initial CCGWAS should be large, in order to limit the number of SNPs that require follow-up (Figure 5). Even larger n would be recommended for the model of random genetic effects (unreported data).

We have extended the simulation methods to compare 1- and 2-stage designs. For a fixed effects model with $M = 1$ and $\beta = \log(1.2)$, a 1-stage design with $n = 8000$ cases and controls has DP $= 0.948$ with $T = 10$. If instead $T = 25\,000$ SNPs are selected at stage 1 based on $n = 1000$ cases and controls, then among the top 10 SNPs in a second stage based on $n = 7000$ independent cases and controls, the DP $= 0.677$. For a random effects model with $\tau = \log(1.2)/0.798$, the corresponding DP estimates are 0.596 for the 1-stage design and 0.485 for the 2-stage design. Thus, the 2-stage design with $n = 1000$ cases and controls in the first stage can lead to an appreciable loss in DP.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute.

REFERENCES

APTEC SYSTEMS. (2005). *The Gauss System, Version 6*. Maple Valley, WA: Aptec Systems, Inc.

ARMITAGE, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386.

BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**, 289–300.

CLAYTON, D. G., WALKER, N. M., SMYTH, D. J., PASK, R., COOPER, J. D., MAIER, L. M., SMINK, L. J., LAM, A. C., OVINGTON, N. R., STEVENS, H. E. *and others* (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics* **37**, 1243–1246.

DEVLIN, B. AND ROEDER, K. (1999). Genomic control for association studies. *Biometrics* **55**, 997–1004.

GAIL, M. H., WEISS, G. H., MANTEL, N. AND OBRIEN, S. J. (1979). Solution to the generalized birthday problem with application to allozyme screening for cell-culture contamination. *Journal of Applied Probability* **16**, 242–251.

PEPE, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford University Press.

PFEIFFER, R. M. AND GAIL, M. H. (2003). Sample size calculations for population- and family-based case-control association studies on marker genotypes. *Genetic Epidemiology* **25**, 136–148.

PRENTICE, R. L. AND PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411.

SASIENI, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261.

SATAGOPAN, J. M., VENKATRAMAN, E. S. AND BEGG, C. B. (2004). Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* **60**, 589–597.

SKOL, A. D., SCOTT, L. J., ABECASIS, G. R. AND BOEHNKE, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics* **38**, 209–213.

VECCHIO, T. J. (1966). Predictive value of a single diagnostic test in unselected populations. *New England Journal of Medicine* **274**, 1171–1173.

WACHOLDER, S., CHANOCK, S., GARCIA-CLOSAS, M., EL GHORMLI, L. AND ROTHMAN, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *Journal of the National Cancer Institute* **96**, 434–442.

YEAGER, M., ORR, N., HAYES, R. B., JACOBS, K. B., KRAFT, P., WACHOLDER, S., MINICHIELLO, M. J., FEARNHEAD, P., YU, K., CHATTERJEE, N. *and others* (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genetics* **39**, 645–649.

ZAYKIN, D. V. AND ZHIVOTOVSKY, L. A. (2005). Ranks of genuine associations in whole-genome scans. *Genetics* **171**, 813–823.

[*Received May 1, 2007; revised July 23, 2007; accepted for publication July 27, 2007*]