Efficient Two-Stage Genome-Wide Association Designs Based on False Positive Report Probabilities

Peter Kraft

EFFICIENT TWO-STAGE GENOME-WIDE ASSOCIATION DESIGNS BASED ON FALSE POSITIVE REPORT PROBABILITIES

PETER KRAFT

Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, 655 Huntington Avenue, Boston , MA 02112, United States of America

Despite recent advances, very-high-throughput (VHT) technologies capable of genotyping hundreds of thousands of SNPs in individual samples remain prohibitively expensive for the large studies necessary to screen substantial sections of the genome for variants with modest effects on disease risk. This paper presents a two-stage strategy, where a portion of available samples are genotyped with VHT technology, and a small number of the most promising variants are genotyped with standard high-throughput techniques in the remaining samples as an independent replication study. The sample sizes in the first and second stages and the corresponding significance levels are chosen to limit False Positive Report Probability (FPRP), while maximizing the number of Expected True Positives (ETPs). (The FPRP is the conditional probability that a marker is not truly associated with disease, given the a significant test for disease-marker association.) For a fixed budget, the two-stage strategy has greater power (a larger number of ETPs) than the single-stage strategy (where all subjects are genotyped using expensive VHT technology). Furthermore, concentrating on the FPRP leads to considerable savings relative to strategies designed to control the family-wise error (e.g. Bonferonni correction). The FPRP and number of ETPs can also accommodate researchers' prior beliefs about the number of causal loci and the magnitude of their effects. The expected number of false positives does not change if the true number and effects of causal loci differs from the specified prior (although the false discovery rate will vary), thus limiting the absolute amount of resources spent chasing "false leads."

1. Introduction

Genome-wide linkage scans have successfully located the genes underlying simple Mendelian disorders, including rare, high-risk hereditary forms of cancer. However, studies based on genetic cosegregation have been less successful in finding susceptibility loci for complex disease; for example, high-risk cancer genes only account small percentage of the familial aggregation of cancer.[*1, 2*] Genome-wide association (GWA) studies are likely to have more power to detect the common, low- to moderate-risk genes that have the greatest impact on morbidity and mortality due to complex disease at the population level.

Advances in our knowledge of the architecture of the human genome—e.g. studies that examine linkage disequilibrium (LD) patterns among dense sets of Single Nucleotide Polymorphisms (SNPs)[*3, 4*]—and advances in high-

throughput genotyping technology have made GWA studies feasible. Several GWA studies are currently underway, including the NCI's Cancer Genetic Markers of Susceptibility study, which aims to identify susceptibility genes for breast and prostate cancer using a series of nested case-control studies.[5]

GWA studies still face a number of design and analysis challenges. Even using the fine-scale correlation structure of the genome to choose a subset of maximally informative SNPs, theoretical and empirical studies suggest hundreds of thousands of SNPs will be needed to cover the genome.[3, 4, 6] Furthermore, to reliably detect low- to moderate-risk genes while controlling the number of false positives, large sample sizes will be necessary.[7-9] Despite rapidly decreasing genotyping costs, over the next few years it will remain prohibitively expensive to genotype all the SNPs needed for a genome-wide scan in all available subjects.

A multi-stage approach may provide a cost-efficient alternative. In the first stage, the full panel of markers is genotyped on a subset of subjects; in the second and subsequent stages, the most promising markers are followed up in the remaining subjects. Given a fixed budget and a fixed sample size, the number of subjects in the first stage and the number of markers to follow up in the second stage can be chosen so as to maximize power while controlling the number of false positives. Satogopan et al. have explored this design in the context of controlling the family-wise error rate (FWER), that is, limiting the number of false positives to zero with high probability.[10-12] This paper introduces a multi-stage framework that limits the False Positive Report Probability (FPRP) recently introduced by Wacholder et al.[13]

The FPRP provides a weaker form of control that is useful in the context of genome-wide association scans. Rather than definitively proving the causality of a locus, the goal of genome-wide association scans will be to suggest a list of candidate genes or regions with high probability of causality for further study. Strong control of the FWER can lead to reduced power or an impractical increase in required sample size. Researchers will be willing to accept a limited number of false positives results if that ensures causal loci will be detected using available resources, especially if any positive results from a genome-wide association scan will be quickly followed up—tested in other populations, studied in vitro, etc.

The next section discusses general technical and logistical constraints on genome wide association studies. It also reviews the FPRP and extends it to two-stage designs. Since the FPRP is a quasi-Bayesian tool, I discuss the choice of priors for genome-wide association scans. The third section presents a hypothetical example involving 100,000 markers and compares the performance of one- and two-stage designs aimed at controlling the FWER and the FPRP.

This example shows that for a fixed budget the two-stage and FPRP approaches can be considerably more powerful (in terms of expected number of causal loci detected) than one-stage and FWER designs. The final discussion includes a comparison of the multi-stage FPRP design and analysis with related designs, such as group sequential sampling and multi-stage designs aimed at controlling the False Discovery Rate.[*14-16*]

2. Materials and methods

For simplicity I assume researchers have access to two classes of genotyping technology: "very high throughput technology" and "high throughput technology." Very high throughput technology (VHT) is good at measuring many genotypes simultaneously on each sample, at a low per-genotype cost. But because of the sheer number of genotypes, each sample is expensive to genotype. Furthermore, the set of SNPs genotyped is relatively inflexible, as developing new arrays or multiplexes is expensive and time consuming. High throughput (HT) technology is somewhat more expensive per sample, but more flexible in terms of choosing the SNPs to be genotyped. It is also currently more widely available.

In the example presented below, I assume that the "very high throughput" technology is used at the first stage of the study, but the more flexible "high throughput" technology is used in subsequent stages. This assumption is not intrinsic to the statistical methods, however, as they simply allow for per-genotype costs to vary across the stages. The distinction between "very high throughput" and "high throughput" technology will likely soon fade, as the former becomes more flexible and start-up costs decline. It may currently be useful in practice to consider three tiers of technology, with per-genotype price breaks occurring between hundreds, tens of thousands, and hundreds of thousands of markers typed per sample.

Notation for various study parameters is presented in Table I. All subsequent discussion is limited to the context of two-stage designs; the calculations should easily extend to designs with three or more stages (where the number of stages is fixed ahead of time).

I assume that researchers have fixed numbers of cases and controls available for study. This would be the case when using DNA samples from existing cohort or case-control studies for genome-wide association. Furthermore, for many rare diseases there is effectively a rather low upper limit on the number of cases that can be studied in a reasonable time frame. I assume that there are an equal number of cases and controls. This assumption is not essential to the method; only minor modifications to power calculations would

be needed to account for case:control ratios different than 1:1. Note that although I focus on case-control studies using unrelated individuals, the design concepts can be easily applied to family-based studies, e.g. where researchers have a fixed number N of case-parent trios. I also assume researchers have a fixed budget B for genotyping.

Table 1. Notation

Parameter	Notation	Example value
Total number of subjects	N	2000*
Number of subjects, stages 1 and 2	N_1, N_2	
Set of markers studied	$\mathcal{M}$	
Total number of marker studied	M	100,000
Allele frequency for marker $m \in \mathcal{M}$	q_m	$\equiv 0.10$
Target FPRP	φ	0.50
Significance threshold at first stage, marker m	α_{1m}	*
Significance threshold at second stage	α_{2m}	*
Possible (non-null) relative risks of disease	$RR_1,...,RR_A$	2,1.5,1.3
Prior probability that a marker has relative risk RR_j	π_j	1,2,4***
Type II error rates at stage i for locus m with RR_j	β_{mji}	**
VHT genotyping cost per subject	K	1
HT:VHT genotyping cost ratio	κ	10

* To be solved for given fixed budget. ** Fixed given N, α_{1m}, α_{2m}. *** Value $\times 10^{-5}$

In the first stage, $N_1 \leq N$ cases and controls are genotyped at M independent markers using very high throughput technology. (I assume these markers are diallelic, but they could also be made up of several correlated SNPs, as would be the case in haplotype-tagging studies.) Each marker m is tested for association with disease; each marker m that is significant at the α_{1m} level is then genotyped using high throughput technology in the remaining N_2 cases and controls. Each marker m that is significantly associated with disease in this second sample at the α_{2m} level is declared "overall significant." (As discussed below, the FPRP depends on both the significance level and the power of the test for association between marker m and disease. As power will depend on the allele frequency q_m which may differ across markers, the significance thresholds at the first and second stage can vary across markers.)

The goal is to find a two-stage design that maximizes the expected number of true associations detected while controlling the FPRP $\leq \varphi$, given the number of subjects N and the budget B. This involves maximizing the expected number of true positives over a grid of designs parameterized by the first stage sample size N_1 and the first stage significance level $\boldsymbol{\alpha}_1 = (\alpha_{11},...,\alpha_{1M})$ (N_2 is fixed given N_1; $\boldsymbol{\alpha}_2$ is fixed given $\boldsymbol{\alpha}_1$, φ, and N_1) such that the expected overall cost remains below B. Assuming that the expected number of truly associated variants is very small relative to the total number M genotyped, the expected cost is

$$\Sigma_m K (N_1 + N2\ \alpha_{1m}\ \kappa),$$

where K is the per-genotype cost of the very high throughput technology and κ is the ratio of high-throughput to very-high-throughput genotyping costs. The calculations developed here could also be used to examine the impact of increasing sample size and budget on the expected number of true positives. This would provide a guide for researchers designing a de novo study or contemplating the cost-benefit ratio for enrolling more subjects or increasing the genotyping budget.

2.1. *Two-stage false positive report probability*

The FPRP is defined as the probability that a variant that has been found statistically significant is actually not associated with disease, but appears statistically significant merely by chance. The FPRP depends on the Type I error rate α of the applied test, the Type II error rate β, and the prior probability that a given locus is truly associated with disease π:

$$\frac{\alpha(1-\pi)}{\alpha(1-\pi)+(1-\beta)\pi}.$$

As originally proposed, the prior density on the strength (relative risk) of the variant-disease association put point masses at unity (no association) and a single non-unity value (which was used to compute the power 1-β).[13, 17] This assumption is easily relaxed, allowing for a range of possible non-null relative risks (and hence a range of βs = $\beta_1,..,\beta_A$) with a range of prior probabilities $\pi_1,\ldots,\pi_A$, leading to the following expression for the FPRP:

$$\frac{\alpha(1-\sum_{j=1\ldots A}\pi_j)}{\alpha(1-\sum_{j=1\ldots A}\pi_j)+\sum_{j=1\ldots A}(1-\beta_j)\pi_j}.$$

To calculate the FPRP for the two-stage design, I assume that only second stage subjects are used to test the most promising markers, so that the tests in the first and second stage are independent. (This approach differs from that of Satogopan et al., who use first and second stage subjects to test the most promising markers.[12] Although limiting second-stage tests to the second-stage sample may reduce power somewhat, it may also be most appropriate if the second-stage sample is a separate study.) The two-stage FPRP for marker m then has a simple form:

$$\frac{\alpha_{2m}\alpha_{1m}(1-\sum\pi_j)}{\alpha_{2m}\alpha_{1m}(1-\sum\pi_j)+\sum(1-\beta_{mj2})(1-\beta_{mj1})\pi_j}.$$

Here the Type II error probabilities β_{mjk} are indexed by the marker $m\in\mathcal{M}$, the relative risks $j=1,\ldots,A$, and stage k; the power 1-β_{mjk} depends on marker allele frequency, relative risk, number of subjects in stage k. Note this expression

assumes the priors $\pi_1,\ldots,\pi_A$ are identical for all markers; this could be easily modified to incorporate prior beliefs about probability a particular SNP plays a causal role. For example, non-synonymous coding SNPs could be upweighted relative to intergenic "tag SNPs."

Assuming further that the priors and tests at multiple loci are independent, the expected number of true positives in a two-stage study can be calculated as:

$$\sum_{m,j}(1-\beta_{mj2})(1-\beta_{mj1})\,\pi_j\,.$$

This expression is used to solve for first stage sample size and first-and second stage significance level that maximize the ETP while controlling the minimum FPRP and remaining under budget.

2.2. *The FPRP and marker choice*

Marker choice is a key factor in genome-wide association studies. Several overlapping paradigms have been suggested.[*4, 8, 18*] An advantage of the FPRP is that the prior probability that a given marker is associated with disease can account for the fact that the set $\mathcal{M}$ of markers measured does not contain all causal loci, and some causal loci may not even be in linkage disequilibrium with markers in $\mathcal{M}$.

2.3. *The FPRP and prior choice*

Choosing a prior for a given candidate gene can be quite difficult and rather subjective. Often the best researchers can do is set a range of priors that spans several orders of magnitude.[*13*] On the other hand, priors on the number of loci with a detectable marginal effect on a disease (and to a lesser extent priors on the sizes of those effects) are somewhat easier to specify, as the number of such loci is believed to be quite small relative to the number of loci screened. At most there will be several score of such loci; perhaps more realistically, there will be less than a dozen. For example, several authors have argued that the prior probability a randomly chosen marker is associated with disease should be on the order of 1 in 10,000.[*13, 19*] For the example presented below, I assume prior probabilities that a marker has a genetic relative risk of 2.0, 1.5 or 1.3 are 1, 2 and 4 in 100,000, respectively. Thus on average seven of the 100,000 markers tested will be truly associated with disease. More realistic priors could be developed using what is known about the distribution of the size of genetic effects in general, characteristics of the disease under study (such as sibling relative risks), and plausible distributions for the allele frequencies of susceptibility loci.[*9*]

3. Results

I calculated the expected number of true positives given a fixed budget for four designs: a single stage case-control study that aims to maximize the number of expected true positives (ETP) while holding the FWER below 5%; a two-stage study that maximizes ETP while also holding the FWER below 5%; a one-stage study that aims to maximize the number of ETP while holding the minimum FPRP below 50%; and a two-stage study that maximizes ETP while holding the minimum FPRP below 50%. Parameter values underlying these calculations are summarized in Table 1. Power was calculated for the standard Pearson's chi-squared statistic for 2×3 tables assuming the risk allele had a multiplicative effect on the risk of disease. For the two-stage studies aimed at controlling the overall FWER, the first- and second-stage samples are analyzed independently, so that the overall Type I error rate is $\alpha_1\alpha_2$. The FWER for one- and two-stage studies is controlled by ensuring the overall Type I error is below $1-(1-\alpha_*)^{1/M}$, where α_* is the target family-wise error rate (e.g. 5%).

Figure 1 shows the maximum expected true positives for the four designs for a range of budgets. The two stage designs are always more powerful than the analogous one-stage designs, although for large enough budgets the one- and two-stage designs have equivalent power. This reflects the fact that if we could afford to genotype all available subjects using the very high-throughput technology, we would. The power advantage of the two-stage designs comes from the ability to genotype more subjects at the second stage; simply splitting the sample and testing the same set of markers in each sub-sample always results in less power. [20] Note that if there were an unlimited number of cases available for enrollment, the two-stage designs would remain more powerful than the one-stage designs as budget increased.

The designs that control the FPRP also have greater power than the analogous designs that control the family-wise error. The expected number of false positives for the FPRP-based designs is larger, as in this case (FPRP=50%) the expected number of false positives for the FPRP is equal to the expected number of false negatives, while for FWER designs it is fixed at ≈ 0.05. However, the expected number of false positives is fixed by design, regardless of the true number of associated markers and their relative risks. This is because limiting the FPRP at φ requires that on average the number of false positives should not exceed $[\varphi / (1-\varphi)] \times$ the expected number of true positives. Thus, when using an FPRP-based design the expected resources spent following false leads is limited, while the chance of detecting a true association is increased.

Figure 2 shows the number of expected true positives when $N_1 = N_2 = 1000$ as a function of α_1, the first stage significance level and roughly equal to the

proportion of markers taken to the second stage. These sample sizes were chosen because the number of expected true positives for the two-stage FPRP design begins to plateau when N_1= 1000, at a budget of 1082 (in units of cost to genotype all M markers on one subject using the very-high-throughput technology). The parameters that maximize the expected true positives for that budget are α_1 = 0.0067 (on average roughly 670 markers are taken to the second stage) and α_2=0.0055. The sharp initial increase in Figure 2 suggests that the power of the two stage design is driven in large part by the Type II error rate of the first stage: if a truly associated marker does not make out of the first stage, no association can be found at the second. On the other hand, the eventual slow decline in expected number of true positives with increasing α_1 is due to the increasingly stringent α_2 level necessary to control the two-stage FPRP.

Finally, Figure 3 shows the number of expected true positives for two-stage designs as a function of α_1 and N_1 for the FPRP and FWER designs. The power surface has similar shape for both designs, although the precise allocation of samples and number of markers to carry to the second stage that maximizes power differs somewhat. For fixed N_1, the FPRP is maximized at higher α_1 than the FWER; for fixed α_1, the FPRP is maximized at a higher N_1.
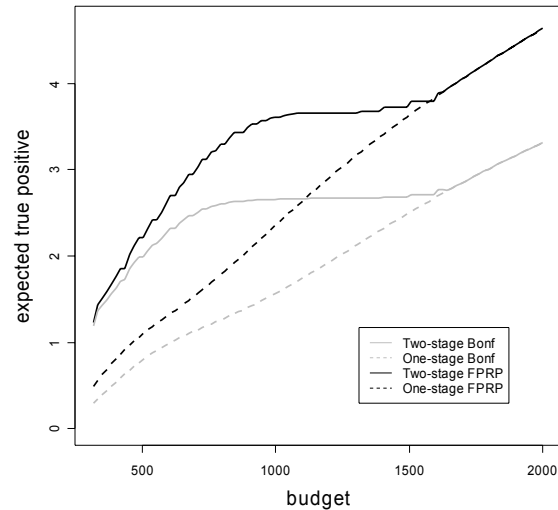

Figure 1. Expected true positives for four study designs as a function of budget, when the total available sample size is fixed at 2000 (see Table 1 for other study parameters). The budget is given in terms of the cost to genotype one subject using the very-high-throughput technology. Thus for a budget of 2000, all subjects could be genotyped using the very-high-throughput technology.

4. Discussion

Although I presented the two-stage FPRP design in the context of case-control studies using unrelated individuals, it can be easily adapted to other contexts, including continuous traits and family-based designs. Other methods for power calculations can also be used. For example, van den Oord and Sullivan proposed a two-stage design very similar to ours, except they used a liability-threshold model in their power calculations.[16] They also assumed marker allele frequencies were randomly distributed similar to Schork[21], although in a genome-wide association analysis using a set of markers chosen from a screening set like the International HapMap it is likely researchers will have accurate estimates of marker allele frequencies beforehand and can use them in power calculations as we did here.

I have also presented this two-stage design in the context of genotyping technologies that allow for subject-specific genotypes, but DNA pooling techniques could also used at either or both stages. This may lead to a further reduction in genotyping costs, although the modest genetic effects anticipated in common complex disease may lie below the signal detection threshold without increasing the number of pools to the point where pooling loses much of its efficiency relative to individual genotyping.

One concern with large genetic association studies using unrelated individuals is that markers may be associated with a trait not because the marker is near a gene on the causal pathway, but because of population stratification or cryptic relatedness.[22-24] In a well-designed study in a relatively homogeneous population such as non-Hispanic European Americans, such effects should be small and could in principle be accounted for using a multinomial analog of the FPRP that calculates the posterior probability of a positive test being a false positive, a true positive association due to population stratification, or a true positive due to linkage disequilibrium with a causal locus. It should be possible to differentiate between the latter two results, because the prior distribution for population stratification effect sizes will differ from that for effect sizes due to linkage disequilibrium.

Sequential designs are an alternative that may be more efficient than the two stage procedure presented here.[14, 15] Instead of treating the sample size at the first stage as fixed, sequential designs consider it to be random and keep adding subjects and recalculating test statistics until one of two significance thresholds is crossed. This allows researchers to stop genotyping early if a significant result is observed or stop early for futility if early returns look "really null." Many such designs are based on controlling the Type I error rate for a single independent variable (here: marker), but it should be straightforward to

adapt them to control the FPRP over multiple markers. Sequential multiple decision procedures[*15*] are intriguingly analogous the FPRP procedure presented here in that instead of controlling the experiment-wide Type I error rate they seek to partition the markers into a set that is enriched for truly associated with disease and a set that is overwhelmingly not associated with disease.
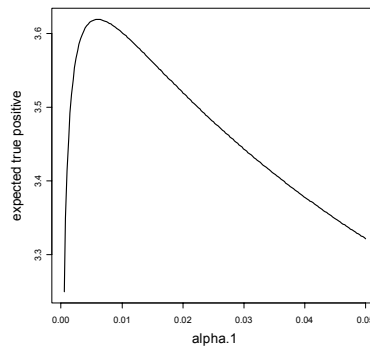

Figure 2. Number of expected true positive results as a function of the first stage significance level α_1 when the sample is split between first and second stage samples, $N_1=N_2=1000$.

However, there are logistical barriers in genetic association studies that may keep such sequential designs from achieving their theoretical gains in efficiency. High- and very-high throughput genotyping technologies require certain economies of scale: many subjects must be genotyped simultaneously, often at many markers simultaneously. This may limit the number of sequential tests. Two- or three-stage studies are likely feasible, but a ten-stage study may be too complex logistically. In particular, it may be infeasible or overly costly to change the set of markers typed more than two or three times. Going from 100,000 markers to ca. 700 is a sufficiently large decrease to justify redesigning genotyping protocols; going from 100,000 to 65,000 may not be.

The False Positive Report Probability is closely related to the False Discovery Rate, defined as the expected fraction of positive tests that are false positives. Procedures that control the FDR provide a weaker form of false-positive control than those which control the family-wise error. The FDR is widely used in hypothesis-generating microarray studies, and its use has been proposed in the context of genetic association studies where researchers are testing many hypotheses.[*16, 25*] When the number of true positives is small relative to the total number of tests (as it is here), the FDR $\approx$ FPRP.[*26*] Although we do not claim the FPRP procedure presented here controls the false

discovery rate, it should be accurate when the number of truly associated markers with clinically relevant effect sizes is very small relative to the total number tested, and the power to detect the truly associated markers is high. Further, the FPRP has a practical advantage over standard FDR procedures that require all test p-values to be ranked: if markers are genotyped in batches, researchers can analyze each batch as it comes in, rather than waiting for data on all markers before moving on.
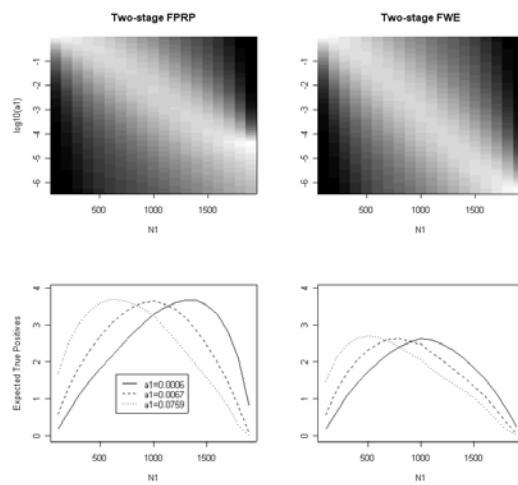

Figure 3. Number of expected true positive results as a function of the first stage significance level α_1 and first stage sample size N_1. Light areas on top two plots correspond to higher numbers of expected true positives. Bottom plots are profiles of these functions for selected values of α_1.

In summary: two-stage designs that limit the number of markers studied in all available subjects can lead to considerable savings. Furthermore, designs that control the two-stage False Positive Report Probability can be more powerful than designs that control the family-wise error rate. The expected number of false positives is higher using FPRP designs, but this should be acceptable in the context of genome-wide association studies, where any positive results would be followed up in other epidemiological and laboratory studies. Although the choice of priors for the FPRP is subjective, this is true of all power calculations, and the quasi-Bayesian framework of the FPRP allows uncertainty in the number of causal loci and their strength to be built in to design calculations.

Acknowledgments

References

1.   A. Balmain, J. Gray, B. Ponder, *Nat Genet* **33 Suppl**, 238 (Mar, 2003).
2.   P. D. Pharoah *et al.*, *Nat Genet* **31**, 33 (May, 2002).
3.   D. A. Hinds *et al.*, *Science* **307**, 1072 (Feb 18, 2005).
4.   The International HapMap Consortium *Nature* **426**, 789 (2003).
5.   *NCI Cancer Bulletin*. (2005), vol. 2, pp. 7.
6.   C. S. Carlson *et al.*, *Nat Genet* **33**, 518 (Apr, 2003).
7.   N. Risch, K. Merikangas, *Science* **273**, 1616 (1996).
8.   J. N. Hirschhorn, M. J. Daly, *Nat Rev Genet* **6**, 95 (Feb, 2005).
9.   W. Y. Wang, B. J. Barratt, D. G. Clayton, J. A. Todd, *Nat Rev Genet* **6**, 109 (Feb, 2005).
10.   J. M. Satagopan, R. C. Elston, *Genet Epidemiol* **25**, 149 (Sep, 2003).
11.   J. M. Satagopan, E. S. Venkatraman, C. B. Begg, *Biometrics* **60**, 589 (Sep, 2004).
12.   J. M. Satagopan, D. A. Verbel, E. S. Venkatraman, K. E. Offit, C. B. Begg, *Biometrics* **58**, 163 (Mar, 2002).
13.   S. Wacholder, S. Chanock, M. Garcia-Closas, L. El Ghormli, N. Rothman, *J Natl Cancer Inst* **96**, 434 (Mar 17, 2004).
14.   I. R. Konig, A. Ziegler, *Hum Hered* **56**, 63 (2003).
15.   M. A. Province, *Genet Epidemiol* **19**, 301 (Dec, 2000).
16.   E. J. van den Oord, P. F. Sullivan, *Hum Hered* **56**, 188 (2003).
17.   D. C. Thomas, D. G. Clayton, *J Natl Cancer Inst* **96**, 421 (Mar 17, 2004).
18.   D. Botstein, N. Risch, *Nat Genet* **33 Suppl**, 228 (Mar, 2003).
19.   J. N. Hirschhorn, D. Altshuler, *J Clin Endocrinol Metab* **87**, 4438 (Oct, 2002).
20.   D. C. Thomas *et al.*, *Am J Epidemiol* **122**, 1080 (1985).
21.   N. Schork, B. Thiel, P. St. Jean, *J Exper Zool* **282**, 133 (1999).
22.   M. L. Freedman *et al.*, *Nat Genet* **36**, 388 (Apr, 2004).
23.   D. Thomas, J. Witte, *Cancer Epidemiol Prev Biom* **11**, 505 (2002).
24.   S. Wacholder, N. Rothman, N. Caporaso, *Cancer Epidemiol Prev Biomarkers* **11**, 513 (2002).
25.   C. Sabatti, S. Service, N. Freimer, *Genetics* **164**, 829 (Jun, 2003).
26.   J. D. Storey, R. Tibshirani, *Proc Natl Acad Sci U S A* **100**, 9440 (Aug 5, 2003).