Multistage Sampling for Genetic Studies*

Robert C. Elston,[1] Danyu Lin,[2] and Gang Zheng[3,**]

[1]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio 44106; email: rce@darwin.case.edu

[2]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599; email: lin@bios.unc.edu

[3]Office of Biostatistics Research, National Heart, Lung and Blood Institute, Bethesda, Maryland 20892; email: zhengg@nhlbi.nih.gov

Annu. Rev. Genomics Hum. Genet. 2007. 8:327–42

First published online as a Review in Advance on May 22, 2007.

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

This article's doi: 10.1146/annurev.genom.8.080706.092357

1527-8204/07/0922-0327$20.00

*The U.S. Government has the right to retain a nonexclusive, royalty-free license in and to any copyright covering this paper.

**The authors are ordered alphabetically.

Key Words

association, DNA pooling, linkage analysis, genome-wide, segregation analysis, self-replication

Abstract

In the past, to study Mendelian diseases, segregating families have been carefully ascertained for segregation analysis, followed by collecting extended multiplex families for linkage analysis. This would then be followed by association studies, using independent case-control samples and/or additional family data. Recently, for complex diseases, the initial sampling has been for a genome-wide linkage analysis, often using independent sib-pairs or nuclear families, to identify candidate regions for follow-up with association studies, again using case-control samples and/or additional family data. We now have the ability to conduct genome-wide association studies using 100,000–500,000 diallelic genetic markers. For such studies we focus especially on efficient two-stage association sampling designs, which can retain nearly optimal statistical power at about half the genotyping cost. Similarly, beginning an association study by genotyping pooled samples may also be a viable option if the cost of accurately pooling DNA samples outweighs genotyping costs. Finally, we note that the sampling of family data for linkage analysis is not a practice that should be automatically discontinued.

INTRODUCTION

In the past 25 years, new methods for genetic epidemiology have been mainly focused on answering whether or not a trait shows Mendelian segregation, cosegregation in families with a genetic marker, and finally association in the population with a genetic marker (10). Segregation analysis has been used to determine an appropriate genetic model for model-based linkage analysis. Segregation analysis has been an efficient initial approach for simple Mendelian diseases, but rarely successful for a complex disease because, when the mode of inheritance requires numerous parameters to be estimated, the estimates have typically been imprecise. Linkage analysis can be done assuming a known genetic model for the trait being studied or in a model-free manner, which may initally be more appropriate for a complex trait. Often, the demonstration of linkage has provided the ultimate proof of a genetic component in the etiology of a disease, whether it is one that segregates in a Mendelian fashion or one that shows a complex pattern of inheritance. As a large number of single nucleotide polymorphisms (SNPs) are now available at relatively low cost, it has been proposed that testing for association between a disease and a large panel of SNPs on unrelated individuals is a more efficient design for gene discovery when gene effects are small. However, association can arise as a result of chance, of population stratification/heterogeneity, of very close linkage, or of pleiotropy. The essential problem for association analysis is to distinguish the first two of these causes from the second two.

The more recent designs and methods of analysis for linkage and association studies have been driven by the advances of molecular technology, which provide more genetic markers at ever-reduced cost. In 1980, Botstein et al. (2) proposed the use of restriction fragment length polymorphisms to find cosegregation of a trait of interest to any point on the genome. Later, microsatellites (63, 64, 68) became the genetic markers used for a genome-wide linkage study, usually numbering about 400, roughly equally spaced across the genome. Today, SNPs are the genetic markers of choice for either linkage or association studies. They are mostly diallelic and abundant, current platforms covering with good potential power about 80% of the Caucasian genome. With the HapMap project (1), millions of SNPs have been identified throughout the genome. In fact, genome-wide association studies with 500,000 SNPs are already being published.

However, there are challenges in performing large-scale association studies to investigate the genetics of complex diseases (3). First, the cost of genotyping is still high. For a genome-wide association study with 500,000 SNPs and a couple of thousand individuals, the genotyping cost alone lies between a quarter and half a million dollars. Second, a large-scale genetic study has a potentially high false positive rate and it is often difficult to replicate results in later independent studies. Multistage designs incorporating both linkage and association components, or cost-efficient association designs with multiple stages of genotyping, are essential to maintain statistical power at an affordable cost with limited false positive rates. In this article, we review the classical paradigm of multistage sampling designs, using both families and unrelated individuals, that have been successful for Mendelian diseases; a two-stage linkage-association design for more complex diseases; and several cost-efficient multistage designs for large association studies. We conclude by noting that there are advantages to collecting family data and conducting a linkage analysis prior to any genome-wide association study.

CLASSICAL PARADIGM

The classical paradigm for genetic analysis in humans has been a multistage procedure. First, after showing that a trait is familial, segregating families have been collected, segregation analysis performed to

verify Mendelian segregation and define the best segregating model for a model-based linkage analysis, and then, after such a linkage analysis, association studies undertaken. This procedure involves multistage sampling of families and individuals. How to ascertain segregating families depends on the disease. For an autosomal dominant disease, families may be ascertained through the parents' phenotypes. For a rare autosomal recessive disease, heterozygote × heterozygote matings are the informative mating types that are most commonly available; however, because the phenotypes of individuals carrying one risk allele and no risk allele are indistinguishable, these mating types can only be ascertained through the presence of an affected offspring (45).

Following a segregation analysis, a model-based linkage study is conducted with marker loci using a sample of multiplex family data. During meiosis, owing to the crossing over of a pair of chromosomes, a parent may transmit a recombinant gamete or a nonrecombinant gamete to an offspring. A genetic marker is linked to a disease locus if the fraction of gametes recombining between them is less than one half. The recombination fraction measures the genetic distance between two loci and a model-based linkage analysis uses family data to test whether or not the recombination fraction is one half. Usually, three or more generations are used to estimate the fraction of gametes that are recombinant. The base 10 logarithm of a likelihood ratio (LOD), often misinterpreted as the logarithm of the odds for linkage (6), is commonly used to test the null hypothesis of no linkage and we reject the null hypothesis when the recombination fraction is significantly less than one half. For genome-wide linkage analysis, in a large sample and the simplest of situations, a maximum LOD score of 3 corresponds to a pointwise P-value of $\sim 10^{-4}$ (6) and indicates a good linkage signal. In other more general situations, a LOD of 3 or more may have much less statistical significance, which has led to the recommendation that LODs be abandoned in favor of P-values (7, 29). Multipoint linkage analysis simultaneously uses the information from all the markers on a chromosome to find positions where trait loci may lie. Here again LODs are usually reported, but their significance is not always clear (69).

When linkage signals are detected, further studies for allelic association due to linkage disequilibrium (LD) can focus on these chromosomal regions. An association study to test LD is complementary to linkage analysis. LD between two loci causes the alleles at these two loci to be associated in the population. Testing for association between the alleles at two loci is most easily done on a random population sample, but both family data and case-control samples can be used.

The study of complex diseases, e.g., diseases caused by segregation at multiple loci with incomplete penetrance, has typically started with linkage analysis followed by association analysis. Linkage analysis can be performed using a model-free method on a sample of family units as small as independent sib-pairs. For example, linkage can be tested using concordantly affected sib-pairs (ASPs), where families with an ASP are ascertained and genotyped to determine the number of alleles that each sib-pair has identical by descent (IBD) at any genomic location; more generally, other affected pairs of relatives may also be studied (30). A sib-pair can share 0, 1, or 2 alleles IBD at a locus, most other relative pairs share 0 or 1 allele IBD. For some mating types, IBD cannot be uniquely determined, in which case the expected IBD sharing is estimated using all the available marker data.

For a quantitative trait locus (QTL), the Haseman-Elston (HE) regression model can be used. The original HE method uses independent sib-pairs to test for linkage of a QTL. The squared measured trait difference between sibs is regressed on the proportion of alleles the sib-pair shares IBD, and testing for linkage is equivalent to testing that this regression model has a negative slope (17). The HE model has been extended to many situations, e.g., larger sib-ships (49) and

Multiplex family: a family containing multiple individuals affected with a disease

LOD: the logarithm (usually to base 10) of a ratio of two likelihoods, that of linkage to that of free recombination

Linkage disequilibrium (LD): population association between the alleles at linked loci

Identical by descent (IBD): alleles at a locus inherited from the same chromosome of an ancestor

Table 1 Parental transmission data for a diallelic SNP

Not transmitted	Transmitted		Total
	A	*B*	
A	a	b	a + b
B	c	d	c + d
Total	a + c	b + d	n

Table 2 Genotype data for a diallelic SNP in a case-control design

	Genotypes			Total
	AA	*AB*	*BB*	
Case	R_0	R_1	R_2	R
Control	S_0	S_1	S_2	S
Total	N_0	N_1	N_2	N

extended pedigrees (60). Provided unaffected sibs are included in the sample, binary affected/unaffected traits can also be analyzed as quantitative data by this same method (9).

If a candidate gene is already suspected, linkage to it can be tested from a sample of parents and an affected offspring (trios) using the transmission/disequilibrium test (TDT) (54). The original TDT compares the frequency that a particular allele is transmitted to an affected offspring from a heterozygous parent to the frequency that it is not so transmitted. **Table 1** presents the parental transmission data for a diallelic marker with alleles *A* and *B*. In *n* trios, *b* heterozygous parents transmitted allele *B* and *c* heterozygous parents transmitted allele *A*. The test statistic, $\text{TDT} = (\text{b} - \text{c})^2/(\text{b} + \text{c})$, has a large sample chi-square distribution with one degree of freedom either when the marker and disease locus are not linked or when they are not in LD. The TDT was proposed as a test of linkage, but it can also be used to test for LD, or both LD and linkage, depending on assumptions and the sampling scheme (47, 53). Whereas it can always be used validly to test for linkage, it is only a valid test for LD when independent trios are sampled or the test properly allows for the nonindependence of sibs. In linkage analysis, several sampling designs may be applied together. For example, if one has both ASPs and parent-offspring trios, then linkage can be tested by testing IBD patterns in the ASPs and by using the TDT for the trios.

After candidate genes have been detected by linkage analysis, we can conduct an association study focusing on these candidate genes with densely spaced SNPs that is either family based or population based. The simplest sampling scheme for family data comprises trios (**Table 1**) for TDT analysis. Schaid & Sommer (42) considered more general score tests for candidate gene association using trios, and the TDT is a special case of these score tests. Note that the TDT and similar tests are family based, i.e., require family data, and are also designed to protect against detecting spurious associations that are due to population stratification, which leads to departure from Hardy-Weinberg equilibrium (HWE) proportions. However, more power is available from the same family-based data if it is not necessary to protect against spurious association (8, 15). In a population-based association study, independent cases and controls are typically sampled. For a SNP with alleles *A* and *B*, the data for a case-control sample are presented in **Table 2** by genotypes and in **Table 3** by alleles. The total sample size is *N*, with *R* cases and *S* controls. Association between the disease and the SNP can be tested by comparing the genotype frequencies (genotype-based analysis) or the allele (*A*, *B*) frequencies (allele-based analysis) (p, q) between cases and controls (13, 37, 51). Let *T* represent all the samples. In **Table 2**, the frequencies of genotypes (*AA*, *AB*, *BB*) in cases and controls are estimated by $(p_0, p_1, p_2) = (R_0/R, R_1/R, R_2/R)$ and $(q_0, q_1, q_2) = (S_0/S,$

Hardy-Weinberg equilibrium (HWE): HWE proportions that remain stable over generations (no selection)

Hardy-Weinberg equilibrium (HWE) proportions: genotypic frequencies that result from random mating

Table 3 Allele-based case-control samples for a diallelic SNP in a case-control design

	Alleles		Total
	A	*B*	
Case	$2R_0 + R_1$	$2R_2 + R_1$	$2R$
Control	$2S_0 + S_1$	$2S_2 + S_1$	$2S$
Total	$2N_0 + N_1$	$2N_2 + N_1$	$2N$

S_1/S, S_2/S), respectively. In **Table 3**, the frequencies for allele A in cases and controls are estimated by $p = (2R_0 + R_1)/(2R)$ and $q = (2S_0 + S_1)/(2S)$, respectively. The trend test statistic and the allele-based test statistic are then respectively given by

$$Z_T(T) = \frac{(RSN)^{1/2}\{(p_2 + p_1/2) - (q_2 + q_1/2)\}}{\{N(N_2 + N_1/2) - (N_2 + N_1/2)^2\}^{1/2}}$$

and

$$Z_A(T) = \frac{(8RSN)^{1/2}(p - q)}{\{(2N_0 + N_1)(2N_2 + N_1)\}^{1/2}}.$$

For the test using $Z_A(T)$, HWE proportions must hold in the population, and then the two tests have similar power. Both test statistics in large samples follow a standard normal distribution if there is no association.

A two-stage linkage-association sampling design has several advantages compared with an association study alone. Linkage analysis can identify candidate gene regions at a discovery stage, so a subsequent association study can focus on candidate genes at a confirmatory stage. If linkage to a genomic region is known to exist, it is less likely that an association to a variant in that region is spurious. Linkage analysis results based on the HE regression model may also be used to help select samples to design an optimal subsequent association study. Wang & Elston (61) define a quantitative linkage score (QLS) that captures the linkage signals; the QLS can guide the selection of subsamples from a linkage study to increase the power of a subsequent association study.

Thus, the classical paradigm has always started with sampling data for segregation analysis, followed by sampling for linkage analysis and then association analysis, the last using case-control data and/or further family data. This multistage method will find multiple loci and multiple variants within each (multiallelic) locus if the disease being studied is due to many rare variants. However, while this approach might be the only method to find rare mutations at a reasonable cost, sampling family data is expensive.

TWO-STAGE DESIGNS FOR ASSOCIATION

Because family data are required to conduct linkage analysis, it might be more efficient to perform a case-control association study without a preceding linkage study when sampling family data is expensive, especially for a large-scale association study (11, 23, 35, 36). Moreover, for late-onset diseases or diseases related to newly developed technology and medical devices, family data are usually not available. For example, in-stent restenosis is a complication related to stent therapy for atherosclerotic coronary artery disease (14). To study any genetic effect on in-stent restenosis, a case-control sample would be far simpler to collect than family data. Case-control samples are also useful for genome-wide association studies with 100,000–500,000 SNPs, because it is much less costly to recruit thousands of unrelated, rather than related, cases and controls.

A cost-effective two-stage design was first developed for case-control association studies in a candidate gene setting. The genotype-based case and control samples for the two-stage design are given in **Table 4**, where the total sample size is $N_{(1)} + N_{(2)}$. Satagopan et al. (39) proposed to genotype all the SNPs under

Table 4 Genotype data for a diallelic SNP, selected for advancement to the second stage, in the first and second stages of a two-stage case-control design

First-stage data T_1				
	Genotypes			
	AA	*AB*	*BB*	Total
Case	R_{01}	R_{11}	R_{21}	$R_{(1)}$
Control	S_{01}	S_{11}	S_{21}	$S_{(1)}$
Total	N_{01}	N_{11}	N_{21}	$N_{(1)}$
Second-stage data T_2				
	Genotypes			
	AA	*AB*	*BB*	Total
Case	R_{02}	R_{12}	R_{22}	$R_{(2)}$
Control	S_{02}	S_{12}	S_{22}	$S_{(2)}$
Total	N_{02}	N_{12}	N_{22}	$N_{(2)}$

study on a fraction, $f_s = R_{(1)}/(R_{(1)} + R_{(2)}) = S_{(1)}/(S_{(1)} + S_{(2)})$, of the cases and controls in the first stage, denoted T_1 (**Table 4**), and test each SNP for association using the test statistic $Z_A(T_1)$. The most significant small fraction of SNPs is then genotyped on the rest of the cases and controls at the second stage, denoted T_2 (**Table 4**). Here again the testing for association uses $Z_A(T)$, but this is now calculated from all the samples. For a powerful two-stage design with a fixed total cost, the general guideline given was to genotype about 25% of the cases and controls ($f_s = 0.25$) in the first stage and then to genotype the 10% ($f_m = 0.10$) most-promising SNPs in the second stage on the rest (75%) of the samples. Given equal cost for the single-stage and two-stage designs, this two-stage approach allows one to genotype 325% more cases and controls than in a single-stage association study ($N_{(1)} + N_{(2)} = 4.25N$), resulting in a substantial increase in power to detect susceptibility SNPs for a disease. In practice, however, the total number of cases and controls may not be arbitrary, but rather prespecified as the sample size of the study. Thus, two-stage designs may be compared with the single-stage design with $N_{(1)} + N_{(2)} = N$.

Satagopan & Elston (38) then modified this two-stage design, assuming that markers are in linkage equilibrium, to find the optimal design parameters that minimize the cost of the study while retaining a prespecified Type I error and power. This approach was further extended (40) to allow for LD among the SNPs, with the result that genotyping 50% of the samples in the first stage and evaluating the most significant 10% of the SNPs in the second stage yields power close to that of the single-stage design. Following the same designs (38–40), Thomas et al. (56) considered estimating relative risks: the risk of having a disease with genotype *AA* relative to that with *BB*, and the risk of having a disease with genotype *AB* relative to that with *BB*—in other words genotype *BB* is used as the reference genotype. When the SNP and the disease locus are not associated (which could result from linkage equilibrium), the RRs are equal to 1. In the second stage they incorporated the data T_1 of the unselected SNPs to enhance power. They demonstrated that a two-stage design with appropriate tag SNP selection can be very cost-efficient compared with the single-stage design.

These two-stage sampling designs and analyses were developed for candidate gene studies or association studies with hundreds, or thousands, of markers. The same designs and analysis strategies can be applied to genome-wide association studies with 100,000–500,000 SNPs. To test a total of $m = 500{,}000$ SNPs, the significance level for each SNP is set to $\alpha = 0.05/m = 1 \times 10^{-7}$. Following (38), Wang et al. (59) studied two-stage designs using allele-based tests and odds ratios for such genome-wide scans. They considered more focused fine mapping in the second stage using the HapMap data and a model with different unit costs of genotyping in the two stages. Let f_s be the fraction of samples genotyped in the first stage and f_m be the fraction of SNPs selected in the first stage for carryover to the second stage in a two-stage design. Let c_1 and c_2 be the per-sample costs of genotyping a SNP at the first and second stages, respectively. The per-sample genotyping cost of the two-stage design is then $c_1 f_s + c_2 (1 - f_s) f_m$, whereas that for a single-stage design is just c_1. If $c_1 > c_2 f_m$, then $c_1 f_s + c_2 (1 - f_s) f_m < c_1$, indicating that a two-stage design reduces the genotyping cost.

A two-stage design is optimal when the design parameters are determined under some optimality criterion, e.g., for minimizing the total cost when the Type I error and power are fixed, or for maximizing the power when the cost and Type I error are fixed. Wang et al. (59) derived optimal configurations for the design parameters, along with sample size and power calculations. Based on 1000 dollars per subject (currently only 250–500 dollars) for 500,000 SNPs in the first stage ($c_1 = 0.2$ cents), and $c_2 = 3.5$ cents in the second stage, they showed that, to minimize the cost, the optimal two-stage design for a range of risk allele

frequencies (0.1–0.9) and realistic odds ratios (1.35–2.0) is to genotype about 30% of all the samples in the first stage at significance levels $\alpha_1 = 0.00362 - 0.00388$, achieving about 90.7% power in the first stage. Then the significance level in the second stage, α_2, should be about 1.6×10^{-7} or 1.7×10^{-7}, compared with the single-stage design with a more stringent level $\alpha = 1 \times 10^{-7}$. To minimize the total sample size under the same allele frequencies and odds ratios, while constraining the cost to be within 110% of the minimum cost, about 40% of the samples are genotyped in the first stage with $\alpha_1 = 0.00510 - 0.00533$, giving 95% power for this stage, and the second-stage level reduces to $\alpha_2 = 1.1 \times 10^{-7}$ or 1.2×10^{-7}, closer to α. In practice, which optimality criterion should be used to design an optimal two-stage study depends on many factors, including disease prevalence and budget.

For the second stage, they proposed a notable extension: to genotype an average of five more densely spaced SNPs, determined from HapMap data, around each SNP significant in the first stage (59). Assuming the same unit costs of genotyping in the two stages, genotyping five more SNPs in the second stage for each selected first-stage SNP may be prohibitive if a large fraction of the SNPs is selected and a small fraction of samples is used in the first stage. In fact, an optimal two-stage design to minimize the total cost, when genotyping five extra SNPs at the second stage for each significant first-stage SNP, entails $\alpha_1 = 0.0005$—much smaller than $\alpha_1 = 0.00362 - 0.00388$ for the same design, but without genotyping the extra SNPs. In addition, about 50% of the samples would be allocated to the second stage, compared with 70% if these extra SNPs were not genotyped. Because 500,000 SNPs may not capture all the genetic variation across the genome, there is a distinct advantage to using more densely spaced SNPs in the second stage to increase coverage of the genome. With this optimal design, the measure of genetic variation covered (see 55) is increased by more than 66% when the extra SNPs are genotyped in the second stage. The disadvantage, however, is that the total cost, although minimized, is almost double that of the optimal two-stage design without genotyping these extra SNPs. When genotyping costs go down even further—currently, only a quarter of what was assumed in Reference 57—genotyping more densely spaced SNPs in the second stage could be promising. However, although one could then genotype extra case-control samples in the second stage for the same cost, and the increased sample size would lead to increased power, it should be noted that, to the extent that genetic variation is not covered by the first-stage SNPs, no amount of extra SNPs at the second stage will help if they are chosen only on the basis of SNPs showing association at the first stage.

Skol et al. (50) studied a simple procedure to combine the test statistics in the two stages under various configurations of the design parameters. When test statistics $Z_A(T_1)$ and $Z_A(T_2)$ are applied to the two stages, respectively, a weighted test statistic for a joint analysis can be written as $Z_\omega(T) = \omega_1 Z_A(T_1) + \omega_2 Z_A(T_2)$, where the weights ω_1 and ω_2 are proportional to the sample sizes used in the two stages. Since the number of markers tested in the second stage is much smaller than the total number of markers, one might be tempted to view stage two as a replication study and assess statistical significance based on the stage-two data alone. However, this strategy almost always results in decreased power to detect genetic association, as compared with analyzing the data from both stages together, despite the fact that less-stringent-significance criteria are used.

It is not obvious how to evaluate statistical significance in two-stage association studies. A common practice is to use the Bonferroni correction based on the total number of markers tested in the first stage (56). This strategy can be highly conservative, especially when the markers are in strong LD, because it assumes that none of the markers eliminated after the first-stage testing would reach statistical

significance if they were genotyped and tested in the second stage, and that the test statistics are independent across all markers. Lin (24) proposed a Monte Carlo procedure that properly accounts for the two-stage sampling and the correlated nature of the SNP data. Through simulation studies, he demonstrated that this procedure provides accurate control of Type I error and can be substantially more powerful than using the Bonferroni correction.

A comparison of various optimal two-stage sampling designs and analyses is summarized in **Table 5**, showing that most studies use no more than 50% of the samples in the first stage and select less than 25% of the SNPs for genotyping in the second stage. In particular, for genome-wide association studies, as the number of SNPs increases, the proportion of SNPs selected for the second stage decreases. One limitation of any two-stage design is that a disease susceptibility SNP may not be selected in the first stage if only a small fraction of SNPs is carried forward to the second stage. For example, Thomas et al. (57) used the level $\alpha_1 = 0.005$ at the first stage for a total of 500,000 SNPs. On average, about 2500 SNPs would then be selected for genotyping and analysis in the second stage. Although the probability that at least one SNP with true association is in the selected 2500 SNPs is unknown, it appears that it may be low (see table 2 in 70). For example, for a genome-wide association study with 200,000 null SNPs and one SNP with true association, the probability that the most significant 20,447 (9931) SNPs would contain the true SNP is 95% when the power for detecting this true SNP is 90% (95%). This transforms to selecting about 5–10% of the total SNPs in the first stage, within the range discussed by Skol et al. (50).

With advances in technology, the genotyping cost is dramatically reduced. When the genotyping cost is of no concern, especially for candidate gene analysis, a single-stage design using all samples in the analysis is more powerful than a two-stage design.

Table 5 Comparisons of two-stage designs and analyses

Comparison items	Satagopan et al. (2002)	Satagopan & Elston (2003)	Satagopan et al. (2004)	Thomas et al. (2004)	Wang et al. (2006)	Skol et al. (2006)
Number of SNPs	Candidate genes	Candidate genes	Candidate genes	Candidate genes	500,000	300,000
Optimality for the design parameters (α_1, α_2, and f)	Maximizing power for given total cost, unrestricted sample size	Minimizing cost for given power	Maximizing power for given total cost, fixed sample size	Fixed design parameters	Various optimality criteria	Fixed design parameters
Statistical procedure	Allele-based analysis	Trend test	Trend test	Estimation of relative risks	Allele-based analysis and odds ratios	Allele-based analysis
Proportion of samples in the first stage	25%	Calculated to achieve 95% power	50%	No	30–40%	10–50%
Proportion of SNPs in the second stage	10%	15–25%	10%	No	0.4–0.5%	1–10%

A single-stage design also allows one to examine multiple phenotypes and to perform multimarker testing. On the other hand, a large-scale case-control association study is costly. Among 500,000 SNPs, only a few of the SNPs are, or are in LD with, true susceptibility SNPs. Genotyping all SNPs using all samples is a waste of resources. With an appropriate choice of design parameters, a two-stage design would reduce genotyping costs up to 50%, yet have power close to the corresponding single-stage design. Therefore, a two-stage sampling design is cost efficient and particularly useful for genome-wide association studies.

In the two-stage designs shown in **Table 5**, the same test statistic is applied to both stages, e.g., the trend test statistic $Z_T(T)$. However, different procedures and test statistics can be applied to the two stages. For example, at the first stage we could pool a fraction of all the DNA samples separately for cases and controls. If the statistic used simply compares the frequencies of an allele in cases and controls, pooling DNA is an efficient way to reduce the genotyping cost in the first stage. Comparing the frequencies of alleles A and B in pooled cases and pooled controls for each SNP is equivalent to the first-stage test of Skol et al. (50). In the second stage, the selected SNPs are separately genotyped for each of the remaining samples. If the genotyping cost in the second stage is much higher than that in the first stage—in particular if more densely spaced SNPs are to be genotyped in the second stage—pooling DNA could be done in the second stage, rather than in the first stage. Pooling DNA in only one stage has the advantage that the individual genotype data can be adjusted for covariates. Pooling DNA has been applied to SNP and haplotype analyses (e.g., 4, 26, 48, 67, 71, 74; for a review, see 46). For quality control, several sets of pools within cases and controls are usually used as replications and, if the pools are not too large, it is possible to obtain good estimates of haplotype frequencies from them (34). However, optimal two-stage designs with pooling DNA require further study and the number of pools is a further design parameter to be determined. As another example, following the same two-stage design as in **Table 4**, we can apply a multimarker analysis to the first stage and select clusters of SNPs for the second stage (25). In this case, the two-stage design is based on clusters of SNPs rather than on single SNPs. This approach is similar to the two-stage design for haplotype analysis proposed by Thomas et al. (57). The optimal configurations of design parameters for two-stage sampling with multimarker analysis are unknown and need further research.

Research on multistage sampling has mainly focused on a binary outcome. For quantitative traits, we can also apply the multistage sampling designs used for binary outcomes by defining two thresholds to determine cases and controls (e.g., for blood pressure, cases could have pressure greater than 160/90 and normal controls pressure less than 120/80). For a QTL, the power to reject the null hypothesis is substantially increased when subjects with more extreme trait values are sampled. Therefore, sampling only the more extreme trait values in each stage may be useful to reduce the cost of genotyping when the cost of measuring trait values is much cheaper than the cost of genotyping.

One disadvantage of case-control association studies is that population stratification or allelic correlation may lead to spurious association (25). Population stratification occurs when subjects with different genetic backgrounds are enrolled. For example, individuals with different ethnic backgrounds may be defined as subpopulations. The subpopulations can cause spurious associations if they have different disease prevalences as well as different allele frequencies for a SNP. An extreme situation of population stratification occurs when cases and controls are sampled from two different ethnic groups. Allelic correlation, arising from cryptic relatedness, may occur within a population if there is not random mating, with the result that HWE does not hold in the population. This can

Hardy-Weinberg disequilibrium (HWD): departure from Hardy-Weinberg equilibrium proportions, often resulting from selection

distort the test statistics, leading to either an underestimate or overestimate of their variances. Adjustments for either population stratification or cryptic relatedness have been studied in the literature for single-stage design, e.g., using structure analysis and a latent-class modeling approach (32, 33, 41, 43, 62), or a genomic control approach (5, 16, 44, 72). The effect of population stratification and allelic correlation on two-stage designs has not been studied and may be more complicated than for single-stage designs. One possibility is to apply genomic control to correct for variance distortion using what are believed to be null SNPs—SNPs that are not in LD with loci underlying the trait under study. Because only a portion of the SNPs are genotyped in the second stage, the candidate null SNPs should be from among those not selected from the first stage for advancement to the second stage. However, genotypes of these null SNPs are only obtained from a fraction of the case-control samples.

MULTISTAGE DESIGNS FOR ASSOCIATION STUDIES

Most research on optimal cost-efficient designs for association studies has focused on two-stage designs. Multistage designs for association can also be considered. Prentice et al. (31) discussed a three-stage design for an association study with 250,000 SNPs. They suggested dividing the case-control samples into three independent portions of equal size for each stage. To allow for an overall average of 2.5 false positives in this three-stage study, the significance level 0.022 was used for each stage. This design has a two-thirds reduction in the genotyping cost, with only 121 SNPs typed at the third stage, each tested at a significance level of 0.022. This design is simple but not optimal, and its power performance is not clear. A similar three-stage screening procedure for genome-wide studies that divides the samples into three fractions, with decreasing sizes from the first stage to the third stage, has also been discussed (19). The significance levels used at the first two stages were liberal, but a more stringent level was proposed for the third stage. The optimality of this design is unknown. Optimal three-stage designs based on different optimality criteria can be obtained, analogous to the optimal two-stage designs. However, because more design parameters are involved, the optimal solutions will be more complicated and difficult to define.

Statistical power can be further enhanced if a "self-replication" analysis stage is employed. This idea of self-replication was introduced by Van Steen et al. (58) for a single-stage family-based association study and was applied to a genome-wide family-based association study with 100,000 SNPs (18). The idea is to identify and use two statistically independent procedures to test each SNP for association. After the first procedure is performed to test for association, the second can be regarded as an independent replication of the results obtained. The main feature of this procedure is that all samples are used in both analysis stages. In a single-stage case-control design, the power of using self-replication is often greater than that of a single-stage analysis without the self-replication (73).

When the two alleles, A and B, of a SNP are independent, HWE holds at this SNP. HWE is often tested by the Hardy-Weinberg disequilibrium (HWD) coefficient (65). Defining $p_1 = R_1/R$, $p_2 = R_2/R$, $q_1 = S_1/S$ and $q_2 = S_2/S$ (**Table 2**), this coefficient is $\Delta_1 = p_2 - (p_2 + p_1/2)^2$ in cases and $\Delta_0 = q_2 - (q_2 + q_1/2)^2$ in controls. Departure from HWE among cases has been used to test for association (12), but has no power under a multiplicative model (28, 66). After standardization, the resulting statistic can be used because it has a large sample chi-squared distribution with 1 degree of freedom under the null hypothesis. HWD has also been used, in both cases and controls, to indicate genotyping errors as a form of quality control for large association studies). Song & Elston (52) proposed using the difference in HWD coefficients between cases and controls, $\Delta_1 - \Delta_0$, to test for association; we refer to this as the Hardy-Weinberg disequilibrium

trend test (HWDTT), denoted $Z_H(T)$. When HWE holds in the population, in large samples $Z_H(T)$ and $Z_T(T)$ are independent under the null hypothesis and so can be used for self-replication. Using all samples, each SNP is first tested by $Z_H(T)$ at the significance level α_1, and those of the m SNPs that are significant at this level are then tested using $Z_T(T)$ at the level α_2, where $\alpha_1\alpha_2 = \alpha/m$ (73).

Using the concept of self-replication, we can introduce into a multisample design the test $Z_H(T)$ as a new analysis stage. For example, in the two-stage design in **Table 4**, we can add this as an extra stage once both sets of data T_1 and T_2 become available: The test statistic $Z_H(T_1 + T_2)$ is calculated on the total sample of $N_{(1)} + N_{(2)}$ individuals and then only those SNPs that are significant (at a predetermined level) are tested using $Z_T(T)$. Thus fewer SNPs are tested, mitigating the multiple testing problem. This self-replication can also be applied to the second stage of any of the two-stage designs in **Table 5** where the trend test is used. Thus, instead of using the trend test in the second stage, we apply the self-replication procedure. In this case, although genotyping costs will not be reduced by introducing a self-replication stage, the power of the two-stage design would be improved (73).

DISCUSSION

In this review, we outlined three types of multistage sampling schemes for genetic studies. The first follows the classic paradigm that has been successfully used for Mendelian diseases: First, familial segregation patterns are examined in families; this is followed by linkage analysis using samples of affected sib-pairs, complete sib-ships, nuclear families, case-parents trios, or large pedigrees; and this is followed by an association study using family and/or individual data.

The second scheme is for more complex diseases: Linkage analysis using family data is first conducted, followed by an association study for candidate genes. This linkage-association design has several advantages. First, results of the prior linkage analysis provide insight into association studies, because markers that are significantly linked to a disease locus can be regarded as providing candidate genes for the association study. Second, testing association between a marker and a disease locus can incorporate results of the linkage analysis to enhance the power of the association study (27, 61). Third, if many rare variants at a single locus can each cause the disease, a linkage study will have more power than an association study to detect that locus. Fourth, a linkage study involves the collection of family data that enables one to detect phenomena such as anticipation and parent-of-origin effects (imprinting).

The third scheme is a multistage design for genome-wide association studies using only case-control samples. This multistage sampling can reduce the genotyping cost while retaining nearly optimal power compared with a single–stage association design. In a two-stage design, a fraction (usually between 25% and 50%) of case-control samples is first genotyped on all 100,000–500,000 SNPs in the first stage, in which association between each SNP and the disease is tested. Only a small fraction of the SNPs (often less than 10%) is selected for genotyping on the rest of the samples. Then the two stages are combined to test the selected SNPs. With high-throughput genotyping technology, this last scheme is most promising for the future, but the advantages of conducting a linkage study prior to a genome-wide association study should not be ignored.

In this review of multistage sampling designs, we focused on designs in which each sample is ascertained conditional on its phenotypes. In pharmarcogenetic or gene-intervention interaction studies of candidate genes, samples are randomized to two treatments (interventions) for a period of time. The genetic characteristics of individuals who develop disease during this period are then compared between the two treatment groups and tested. The genotyping cost may not

then be an issue because, in such a prospective case-control study, follow-up of individuals usually costs much more than genotyping. For example, to study whether or not regular use of an inhaled b2-adrenergic agonist worsens airflow and clinical outcomes in asthma patients, a small-scale retrospective case-control study was carried out in a first stage (22). This suggested that adverse events occurred in individuals with homozygosity for arginine (Arg/Arg), rather than glycine (Gly/Gly), at aminoacid residue 16 of the receptor. No significant events were identified with heterozygosity (Arg/Gly). Therefore, in a second stage, only Gly/Gly and Arg/Arg individuals were identified and enrolled, and then regularly scheduled treatment with albuterol or placebo was given to each matched pair (Gly/Gly-Arg/Arg, matching for forced expiratory volume in 1 s, i.e., FEV1) in a masked, crossover design (21). This two-stage genotype-stratified design dramatically reduced the total cost of the prospective trial compared with the same trial that would have enrolled random samples without selective genotyping.

In conclusion, most genetic studies need to have independent replication, in particular for a large study, because there is a danger of a high false positive rate from testing many SNPs. Once the results of a genotype-phenotype association have been replicated, additional sampling may be needed for fine mapping, resequencing, functional studies, and, finally, developing personalized medicine.

SUMMARY POINTS

1. The classical paradigm for the genetic analysis of a familial disease in humans has been to collect family samples for segregation analysis and linkage analysis, followed by collecting further samples for association studies. This multistage sampling procedure has been extremely successful for Mendelian diseases.
2. For complex diseases, the first stage has tended to be a genome-wide linkage study to suggest candidate regions for further fine-mapping and association studies.
3. A well-designed optimal two-stage association study, possibly pooling DNA at one stage, is cost-efficient for genome-wide association studies. It has power close to that of a single-stage design, but with up to a 50% reduction in the genotyping cost.
4. Three-stage association designs are also promising, but need further study of their optimality.
5. Statistical methods need to be developed to control for confounding factors and to test for gene-gene and gene-environment interactions in multistage association studies.
6. Replication and resequencing are important stages of a sampling design for genetic studies.
7. Collecting a sample of families for a linkage study before conducting a genome-wide association study may still be useful, either to reduce the cost when there is allelic heterogeneity or to detect generational effects.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported in part by grants from the U.S. Public Health Service, resource grant RR03655 from the National Center for Research Resources, research grants GM-28356 from the National Institute of General Medical Sciences and CA82659 from the National Cancer Institute, and Cancer Center Support Grant P30CAD43703 from the National Cancer Institute. We wish to thank Vernon Chinchilli for providing references on the asthma studies.

LITERATURE CITED

1. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P. 2005. A haplotype map of the human genome. *Nature* 437:1299–320
2. Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32:314–31
3. Clark AG, Boerwinkle E, Hixson J, Sing CF. 2005. Determinants of the success of whole-genome association testing. *Genome Res.* 15:1463–67
4. Craig DW, Huentelman MJ, Hu-Lince D, Zismann VL, Kruer MC, et al. 2005. Determinants of the success of whole-genome identification of disease causing loci using an array-based genotyping approach on pooled DNA. *BMC Genomics* 6:138
5. Devlin B, Roeder K. 1999. Genomic control for association studies. *Biometrics* 55:997–1004
6. Elston RC. 1994. P values, power, and pitfalls in the linkage analysis of psychiatric disorders. In *Genetic Approaches to Mental Disorders*, ed. ES Gershon, CR Cloninger, pp. 3–21. Am. Psych Press, Washington, D.C.
7. Elston RC. 1998. Methods of linkage analysis—and the assumptions underlying them. *Am. J. Hum. Genet.* 63(4):931–34
8. Elston RC, George VT, Severtson F. 1992. The Elston-Stewart algorithm for continuous genotypes and environmental factors. *Hum. Hered.* 41:16–27
9. Elston RC, Song D, Iyengar SK. 2005. Mathematical assumptions versus biological reality: myths in affected sib-pair linkage analysis. *Am. J. Hum. Genet.* 76:152–56
10. Elston RC, Spence MA. 2006. Advances in statistical human genetics over the last 25 years. *Statist. Med.* 25:3049–80
11. Evangelou E, Trikalinos TA, Salanti G, Ioannodos JPA. 2006. Family-based versus unrelated case-control designs for genetic associations. *PLoS Genet.* 2:1147–55
12. Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, et al. 1996. A novel MHC class I-like gene is mutated in patients with hereditary hemochromatosis. *Nat. Genet.* 13(4):399–408
13. Freidlin B, Zheng G, Li Z, Gastwirth JL. 2002. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum. Hered.* 53:146–52
14. Ganesh SK, Skelding KA, Mehta L, O'Neill K, Joo J, et al. 2004. Rationale and study design of the CardioGene Study: genomics of in-stent restenosis. *Pharmacogenomics* 5:949–1004
15. George VT, Elston RC. 1987. Testing the association between polymorphic markers and quantitative traits in pedigrees. *Genet. Epidemiol.* 4:193–202
16. Gorroochurn P, Heiman GA, Hodge SE, Greenberg DA. 2006. Centralization the noncentral chi-square: a new method to correct for population stratification in genetic case-control association studies. *Genet. Epidemiol.* 30:277–89
17. Haseman JK, Elston RC. 1972. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* 2:3–19

18. Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, et al. 2006. A common genetic variant is associated with adult and childhood obesity. *Science* 312:279–83
19. Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6:95–108
20. Hoh J, Wille A, Ott J. 2001. Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res.* 11(12):2115–19
21. Israel E, Chinchilli VM, Ford JG, Boushey HA, Cherniack RM, et al. 2004. Use of regularly scheduled albuterol treatment in asthma: genotype-stratified, randomised, placebo-controlled cross-over trial. *Lancet* 364:1505–12
22. Israel E, Drazen JM, Liggett SB, Boushey HA, Cherniack RM, et al. 2000. The effect of polymorphisms of the beta (2)-adrenergic receptor on the response to regular use of albuterol in asthma. *Am. J. Respir. Crit. Care Med.* 162:75–80
23. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* 308:385–89
24. Lin DY. 2006. Evaluating statistical significance in two-stage genomewide association studies. *Am. J. Hum. Genet.* 78:505–09
25. Marchini J, Cardon LR, Phillips MS, Donnelly P. 2004. The effects of human population structure on large genetic association studies. *Nat. Genet.* 36:512–17
26. Meaburn E, Butcher LM, Liu L, Fernandes C, Hansen V, et al. 2005. Genotyping DNA pools on microarrays: tackling the QTL problem of large samples and large numbers of SNPs. *BMC Genomics* 6:52
27. Nagelkerke NJD, Hoebee B, Teunis P, Kimman TG. 2004. Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. *Eur. J. Hum. Genet.* 12:964–97
28. Nielson DM, Ehm MG, Weir BS. 1998. Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *Am. J. Hum. Genet.* 63:1531–40
29. Nyholt DR. 2000. All LODs are not created equal. *Am. J. Hum. Genet.* 67(2):282–88
30. Olson JM. 1999. A general conditional-logistic model for affected-relative-pair linkage studies. *Am. J. Hum. Genet.* 65:1760–69
31. Prentice RL, Pettinger M, Anderson GL. 2005. Statistical issues arising in the Women's Health Initiative. *Biometrics* 61:899–941
32. Pritchard JK, Rosenberg NA. 1999. Use of unlinked genetic markers to detect population structure in genetic association studies. *Am. J. Hum. Genet.* 65:220–28
33. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2001. Association mapping in structured populations. *Am. J. Hum. Genet.* 67:170–81
34. Quade SRE, Elston RC, Goddard KAB. 2005. Estimating haplotype frequencies in pooled DNA samples when there is genotyping error. *BMC Genet.* 6:25
35. Risch N. 2000. Searching for genetic determinants in the new millennium. *Nature* 405:847–56
36. Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* 273:1516–17
37. Sasieni PD. 1997. From genotypes to genes: doubling the sample size. *Biometrics* 53:1253–61
38. Satagopan JM, Elston RC. 2003. Optimal two-stage genotyping in population-based association studies. *Genet. Epidemiol.* 25:149–57
39. Satagopan JM, Verbel DA, Venkatramanm ES, Offit KE, Begg CB. 2002. Two-stage design for gene-disease association studies. *Biometrics* 58:163–70
40. Satagopan JM, Veukatraman ES, Begg CB. 2004. Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* 60:589–97

41. Sattern GA, Flanders WD, Yang Q. 2001. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am. J. Hum. Genet.* 68:466–77
42. Schaid DJ, Sommer SS. 1993. Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am. J. Hum. Genet.* 53:1114–26
43. Schork NJ, Fallin D, Thiel B, Xu X, Broeckel U, et al. 2001. The future of genetic case-control studies. *Adv. Genet.* 42:191–212
44. Setakis E, Stirnadel H, Balding DJ. 2006. Logistic regression protects against population structure in genetic association studies. *Genome Res.* 16:290–96
45. Sham PC. 1998. *Statistics in Human Genetics*. London: Arnold
46. Sham PC, Bader JS, Craig I, O'Donovan M, Owen M. 2002. DNA pooling: a tool for large-scale association studies. *Nat. Rev. Genet.* 3:862–71
47. Sham PC, Curtis D. 1995. An extended transmission/disequilibrium test (TDT) for multiallele marker loci. *Ann. Hum. Genet.* 59:855–56
48. Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A. 1998. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Res.* 8:111–23
49. Shete S, Jacobs KB, Elston RC. 2003. Adding further power to the Haseman and Elston method for detecting linkage in larger sibships: weighting sums and differences. *Hum. Hered.* 55:79–85
50. Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2006. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* 38:209–13
51. Slager SL, Schaid DJ. 2001. Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. *Hum. Hered.* 52:149–53
52. Song KJ, Elston RC. 2006. A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. *Statist. Med.* 25:105–26
53. Spielman RS, McGinnis RE, Ewens WJ. 1994. The transmission/disequilibrium test detects cosegregation and linkage. *Am. J. Hum. Genet.* 54:559–60
54. Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium—the insulin gene region and insulin-dependent diabetes-mellitus (IDDM). *Am. J. Hum. Genet.* 52:506–16
55. Stram DO. 2004. Tag SNP selection for association studies. *Genet. Epidemiol.* 27:365–74
56. Thomas DC, Haile RW, Duggan D. 2005. Recent developments in genomewide association scans: a workshop summary and review. *Am. J. Hum. Genet.* 77:337–45
57. Thomas DC, Xie R, Gebregziabher M. 2004. Two-stage sampling designs for gene association studies. *Genet. Epidemiol.* 27:401–14
58. Van Steen K, McQueen MB, Herbert A, Raby B, Lyon H, et al. 2005. Genomic screening and replication using the same data set in family-based association testing. *Nat. Genet.* 37:683–91
59. Wang H, Thomas DC, Pe'er I, Stram DO. 2006. Optimal two-stage genotyping designs for genome-wide association scans. *Genet. Epidemiol.* 30:356–68
60. Wang T, Elston RC. 2005. Two-level Haseman-Elston regression for general pedigree data analysis. *Genet. Epidemiol.* 29:12–22
61. Wang T, Elston RC. 2006. A quantitative linkage score for an association study following a linkage analysis. *BMC Genet.* 7:5
62. Wang Y, Localio R, Rebbeck TR. 2005. Bias correction with a single null marker for population stratification in candidate gene association studies. *Hum. Hered.* 59:165–75

63. Weber JL, May PE. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* 44:388–96
64. Weber JL, Broman KW. 2001. Genotyping for human whole-genome scans: past, present, and future. *Adv. Genet.* 42:77–96
65. Weir BS. 1996. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sunderland, Massachusetts: Sinauer Assoc.
66. Wittke-Thompson JK, Pluzhnikov A, Cox NJ. 2005. Rational inferences about departures from Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* 76:967–86
67. Yang YN, Zhang JS, Hoh J, Matsuda F, Xu P, et al. 2003. Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA. *Proc. Nat. Acad. Sci. USA* 100:7225–30
68. Yuan B, Vaske D, Weber JL, Beck L, Sheffield VC. 1997. Improved set of short-tandem-repeat polymorphisms for screening the human genome. *Am. J. Hum. Genet.* 60:459–60
69. Xing C, Elston RC. 2006. Distribution and magnitude of Type I error of the model-based multipoint lod score. Implications for multipoint mod score. *Genet. Epidemiol.* 30:447–58
70. Zaykin DV, Zhivotovsky LA. 2005. Ranks of genuine associations in whole-genome scans. *Genetics* 171:813–23
71. Zeng D, Lin DY. 2005. Estimating haplotype-disease associations with pooled genotype data. *Genet. Epidemiol.* 28:70–82
72. Zheng G, Freidlin B, Gastwirth JL. 2006. Robust genomic control for association studies. *Am. J. Hum. Genet.* 78:350–56
73. Zheng G, Song KJ, Elston RJ. 2007. Combining trend tests for genome-wide association studies: a double trend test. Submitted for publication
74. Zuo Y, Zou G, Zhao H. 2006. Two-stage designs in case-control association analysis. *Genetics* 173:1747–60

Annual Review of Genomics and Human Genetics

Volume 8, 2007

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* chapters may be found at http://genom.annualreviews.org/