BMC Genomics

Methodology article

Open Access

# Incorporation of genetic model parameters for cost-effective designs of genetic association studies using DNA pooling

Fei Ji[1], Stephen J Finch[2], Chad Haynes[1], Nancy R Mendell[2] and Derek Gordon*[3]

Address: [1]Lab of Statistical Genetics, Rockefeller University, New York, NY, USA, [2]Department of Applied Math and Statistics, Stony Brook University, Stony Brook, NY, USA and [3]Department of Genetics, Rutgers University, Piscataway, NJ, USA

Email: Fei Ji - fji@mail.rockefeller.edu; Stephen J Finch - sjfinch@optonline.net; Chad Haynes - Chad.Haynes@mail.rockefeller.edu; Nancy R Mendell - nmendell@notes.cc.sunysb.edu; Derek Gordon* - gordon@biology.rutgers.edu

* Corresponding author

Published: 16 July 2007

*BMC Genomics* 2007, **8**:238 doi:10.1186/1471-2164-8-238

Received: 19 December 2006
Accepted: 16 July 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/238

## Abstract

**Background:** Studies of association methods using DNA pooling of single nucleotide polymorphisms (SNPs) have focused primarily on the effects of "machine-error", number of replicates, and the size of the pool. We use the non-centrality parameter (NCP) for the analysis of variance test to compute the approximate power for genetic association tests with DNA pooling data on cases and controls. We incorporate genetic model parameters into the computation of the NCP. Parameters involved in the power calculation are disease allele frequency, frequency of the marker SNP allele in coupling with the disease locus, disease prevalence, genotype relative risk, sample size, genetic model, number of pools, number of replicates of each pool, and the proportion of variance of the pooled frequency estimate due to machine variability. We compute power for different settings of number of replicates and total number of genotypings when the genetic model parameters are fixed. Several significance levels are considered, including stringent significance levels (due to the increasing popularity of 100 K and 500 K SNP "chip" data). We use a factorial design with two to four settings of each parameter and multiple regression analysis to assess which parameters most significantly affect power.

**Results:** The power can increase substantially as the genotyping number increases. For a fixed number of genotypings, the power is a function of the number of replicates of each pool such that there is a setting with maximum power. The four most significant parameters affecting power for association are: (1) genotype relative risk, (2) genetic model, (3) sample size, and (4) the interaction term between disease and SNP marker allele probabilities.

**Conclusion:** For a fixed number of genotypings, there is an optimal number of replicates of each pool that increases as the number of genotypings increases. Power is not substantially reduced when the number of replicates is close to but not equal to the optimal setting.

## Background

Case/control genetic association studies are used as a means of localizing susceptibility genes for a complex disease. With the recent development of technologies that can determine the genotypes for hundreds of thousands of single nucleotide polymorphisms (SNPs) across the human genome, such studies are now being reported in the literature [1-3]. Design issues such as power to detect association using these technologies are also being published [4,5]. Since a critical requirement for such studies to be sufficiently powered is that the disequilibrium among the disease allele and neighboring marker alleles be large, marker density needs to be high. If the effect size for a complex disease is small (e.g., genotype relative risks [6] on the order of 1.5 to 2), the sample size required to detect association may be thousands of cases and controls [4,5,7-9]. Therefore, researchers often consider genotyping technologies such as DNA pooling [10-13] as an initial strategy to identify genomic regions that may harbor susceptibility loci in an effort to reduce cost (time and money) (e.g.,[14,15]). Advantages of DNA pooling technologies include (a sometimes substantial) reduction in genotyping cost when performing multi-stage association studies to identify disease susceptibility genes. Potential disadvantages include reliance on a number of assumptions related to statistical design and analysis. For example, a key assumption is that the intensity measure has an expected value equal to the allele frequency. Another potential disadvantage is that DNA pooling techniques may not detect disease mode of inheritances that deviate from dominant or recessive modes. For example, DNA pooling techniques will be underpowered to detect disease genes that operate in an over-dominant form.

Sham et al. reviewed currently available technologies for DNA pooling [10]. The statistical analysis of data from pooled DNA studies uses analysis of variance (ANOVA) procedures that have algorithms for calculating power to detect unequal allele probabilities. A major design issue when using DNA pooling technologies is the measurement error as compared with the gold standard method of individual genotyping.

Research has been done regarding specification of study parameter settings to maximize power [10,16,17]. The research question addressed in this work is: assuming a certain level of measurement error, what settings of study design parameters maximize the power to detect association? More specifically, we study the sensitivity of power to changes in design parameters (e.g., total sample size, differing numbers of genotypings, number of pools, and genetic model parameters). We present a closed form approximation to the power in terms of the genetic model, pooling measurement error model, and the study parameters (e.g., number of pools, number of replicates per pool, sample size) and we perform a systematic study of the design parameters to identify which have the greatest effect on power to detect association for DNA pooling studies.

## Results

The pooled DNA association studies considered here have equal number of cases and controls $N$. For a fixed number of total subjects (cases and controls), an equal number of cases and controls yields maximal power for association [7,8]. The $N$ subjects in each group are randomly assigned to one of $J$ pools, each of size $T$ (so that $N = J \times T$). Each of the $J$ pools has $K$ replicate measures, so that the number of case genotypings is equal to the number of control genotypings ($G = J \times K$). The data analyzed in the study are the estimated allele frequencies $Y_{ijk}$ of the more common allele (called "2"), where the index $i$ is 0 for cases and 1 for controls, the index $j$ ranges from 1 to $J$, and the index $k$ ranges from 1 to $K$. The variance of $Y_{ijk}$ has two components, one due to the sampling variability of the frequency of allele 2 in each pool (denoted by $\sigma_{P,i}^2$ here) and the other due to the variability of the measurement process of the pooled material (denoted by $\sigma_E^2 = \sigma_{P,i}^2(m-1)$ here). We refer to the term $m$ as the measure of the *machine replicability variance factor*. The quality of the estimate of the pooled frequency as measured by its variance is parameterized so that is proportional to the sampling variance of the allele 2 frequency and is assumed to be independent of pool size or other pooling parameters. When the number of pools is $J \geq 2$, the structure of a pooled DNA study is an example of a two-stage nested design [18]. Its statistical analysis is conventionally organized in an ANOVA table as in Table 1, with the null hypothesis that the case allele 2 frequency is equal to the control allele 2 frequency. This hypothesis is tested using the statistic $F = \frac{SS_A / 1}{SS_P / [2(J-1)]} = \frac{MS_A}{MS_P}$. Here, $SS_A$ is the sum of squares of the case/control averages, $SS_P$ is the sum of squares of the pool averages within a group about the group pool mean, and is the basis of an estimate of the variance of a pool average frequency. The term $MS_A$ is the mean square of $SS_A$, which by definition is just $SS_A$ divided by the degrees of freedom (df), and similarly for $MS_P$. Under the null hypothesis, $MS_A$ is also the basis of an estimate of the variance of a pool average frequency. When the null hypothesis is false, on average, $MS_A$ is increased as shown in its expected mean square.

**Table 1: The analysis of variance table for a two-stage nested design**

| | | ANOVA Table | |
|---|---|---|---|
| Source | DF | SS | E(MS) |
| Case or control ($\alpha$) | 1 | $\sum_i \sum_j \sum_k (Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet})^2 = JK\sum_i (Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet})^2$ | $\frac{JK\sum \alpha_i^2}{I-1} + (K\sigma_P^2 + \sigma_E^2)$ |
| Pools nested in case or control ($P$) | 2($J$ - 1) | $\sum_i \sum_j \sum_k (Y_{ij\bullet} - Y_{i\bullet\bullet})^2 = K\sum_i \sum_j (Y_{ij\bullet} - Y_{i\bullet\bullet})^2$ | $(K\sigma_P^2 + \sigma_E^2)$ |
| Replicates ($E$) | $IJ$($K$ - 1) | $\sum_i \sum_j \sum_k (Y_{ijk} - Y_{ij\bullet})^2$ | $\sigma_E^2$ |

Abbreviations for column headings are as noted below.
DF: the degrees of freedom for the respective source row;
SS: the sum of squares for the respective source row;
E(MS): expectation of the mean square for the respective source row.
The sums of squares are based on the following terms:

$$Y_{ij\bullet} = \frac{\sum_{k=1}^{K} Y_{ijk}}{K}, Y_{i\bullet\bullet} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}}{JK}, \text{ and } Y_{\bullet\bullet\bullet} = \frac{\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}}{2JK}.$$

The model we consider for individual pooled allele frequency estimates is

$$Y_{ijk} = \mu + \alpha_i + P_{j(i)} + \sigma_E E_{ijk} = \frac{A_{ijk}}{A_{ijk} + B_{ijk}},$$

where the "group" effect associated with cases or controls is $\alpha_i = E(\Pi_i) - E(\frac{\Pi_0 + \Pi_1}{2})$, $i$ = 0,1 subject to the constraint $\sum\alpha_i = 0$. The random effect associated with the $j$th pool in either cases or controls is $P_{j(i)} \sim N(0, \sigma_{P,i}^2)$, with $\sigma_{P,i}^2 = \frac{J\tau_i^2}{N}$. Finally, {$E_{ijk}$} are independent $N$(0, 1) random variables and $\tau_i^2$ is the variance of the allele frequency in the $i$th group.

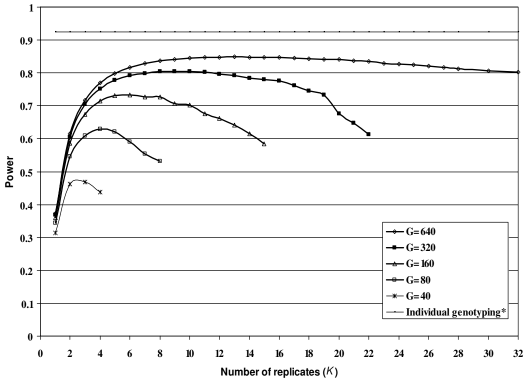
The power calculation of the $F$-test, the standard statistical procedure used when testing allele frequency differences for DNA pooling, requires the non-centrality parameter (NCP) of the test. Its approximate value is given in equation 1 of the Methods and Technical Issues section below. The NCP is a function of the difference between the case and control allele 2 frequencies, the quality of the pooling estimate of these probabilities, the number of cases and controls, the number of replications of DNA measurements of each pool, and the size of each pool.

When the number of replicates $K$ is fixed, the approximate NCP is constant with respect to the number of pools ($J$). When the number of pools $J$ is larger, the denominator degrees of freedom (df) are larger, so that the power of the $F$-test is greater. That is, smaller pool sizes $T = N/J$ for larger $J$, have greater power. The protocol of genotyping each subject has $T$ = 1, which is the most powerful allele frequency testing protocol. That is, if genotype cost is not an issue, it is always most powerful to individually genotype all subjects.

When the total number of genotypings ($G = J \times K$) is fixed, as is the situation for a fixed budget, the optimal choice of $J$ and $K$ is more complex. When one knows the genetic model parameters, one can examine the power using a range of values of $J$ and $K$ (and hence $T$) to find settings with high power. We seek to find $K_o(G)$, the number of replicates that has greatest power when there are $G$ genotypings. For example, Figure 1 is based on a recessive mode of inheritance (MOI) with $N$ = 10,000, prevalence $\phi$ = 0.05, disease allele frequency $p_d$ = 0.15, relative risk of homozygous for disease allele ($R_2$) is 3, linkage disequilibrium $p_r = 0.9R_{\max}^2$, (where $R_{\max}^2$ is the maximum disequilibrium value between the disease allele and the coupling SNP allele; also see PAWE-3D website Helpfile [19], minor SNP allele frequency $q_1 = 0.35$, and machine replicability variance factor $m = 2.25$. We set $G = J \times K$ to 80, 160, 320, and 640 with significance level 0.0001. The power increases substantially as $G$ increases. For example, the maximum power is 0.73 with 80 genotypings when $K_o(80) = 4$; that is, 4 replicates of each of 20 pools. It increases to 0.91 with 640 genotypings when $K_o(640) = 13$; that is 13 replicates each of 49 pools. The power of the chi-squared 2 × 2 test of independence when each subject is individually genotyped is 0.97. With 640 genotypings, the power with $K = 4$ is 0.85. The increase of power from 0.73 to 0.91 is obtained through additional genotyping effort rather than increased sampling of subjects. Also note that the power when $K = 1$ is always substantially less than the power using the optimal choice of $K$; that is, replication of pool measurement is always advantageous.

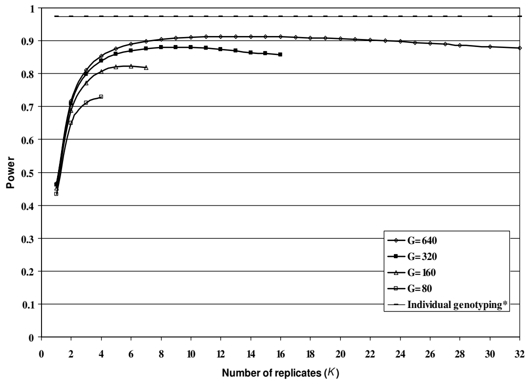
Figure 2 is based on a dominant MOI with $N = 5{,}000$, prevalence $\phi = 0.05$, disease allele frequency $p_d = 0.15$, relative risk of a genotype with at least one copy of the disease allele is 1.5, linkage disequilibrium $p_r = 0.9R^2_{\max}$, minor SNP allele frequency $q_1 = 0.35$, and machine replicability variance factor $m = 2.25$. We set the number of genotypings $J \times K$ to 40, 80, 160, 320, and 640. The pattern is similar to that of Figure 1. A program is available from the corresponding author to produce these numbers for user specified settings.

We note that, although results are not presented, we performed analyses similar to those presented in Figures 1 and 2 for a multiplicative MOI. The conclusions were the same, with results being very similar to the dominant MOI results (Figure 2). We omit these results in the interest of brevity.



## Figure 1

**Power as a function of number of replicates (*K*) for fixed number of genotypings (*G* = *J* × *K*) with recessive mode of inheritance**. Power values presented here are for studies with $N = 10000$, prevalence $\phi = 0.05$, disease allele frequency $p_d = 0.15$, relative risk of homozygous for disease allele $R_2 = 3$, minor SNP marker allele frequency $q_1 = 0.35$, machine replicability variance factor $m = 2.25$, linkage disequilibrium $p_r = 0.9R^2_{\max}$ and significance level alpha = 0.0001. *The horizontal line represents the power for specified parameters with individual genotyping using the 2 × 2 test of independence. Power with individual genotyping was computed using the method implemented in the Power for Association With Error (PAWE) website [27].



## Figure 2

**Power as a function of number of replicates (*K*) for fixed number of genotypings (*G* = *J* × *K*) with dominant mode of inheritance**. Power values presented here are for studies with $N = 5000$, prevalence $\phi = 0.05$, disease allele frequency $p_d = 0.15$, relative risk of a genotype with at least one copy of the disease allele = 1.5, minor SNP marker allele frequency $q_1 = 0.35$, machine replicability variance factor $m = 2.25$, linkage disequilibrium $p_r = 0.9R^2_{\max}$ and significance level alpha = 0.0001. *The horizontal line represents the power for specified parameters with individual genotyping using the 2 × 2 test of independence. Power with individual genotyping was computed using the method implemented in the Power for Association With Error (PAWE) website [27].

The program mentioned above was used to create Table 2, which considers the robustness of design choices when studying a disease with prevalence equal to 0.05. We consider both dominant and recessive MOI with genetic relative risk (GRR) values ranging from 1.5 to 2.2 for specified levels of significance, linkage disequilibrium, sample size, minor SNP marker allele frequency and quality of pooling measurement *m*. We examine the range of numbers of genotypings $J \times K$ between 40 and 640. Table 2 gives the maximum power for each number of genotypings, $K_o(G)$, and the range of *K* settings that produce power within 95% of the maximal power. As in Figures 1 and 2, the most important result is that increasing *G* always substantially increases power. For example, in scenario 1 with 10,000 subjects in each group, recessive MOI, relative risk 2.2, and level of significance 0.0001, the maximal power is 38% with 40 genotypings compared to 77% with 640 genotypings. Similar patterns hold for the other situations considered. The value of $K_o(G)$ increases at a less than linear rate as *G* increases. Typically, the decrease in power associated with using a value of *K* slightly different from $K_o(G)$ is relatively small; that is, the power of the procedure is relatively insensitive to choice of *K*. While $K = 4$ is optimal or close to optimal when the number of genotypings is small (i.e. $G = 40$ or 80), $K_o(G)$ increases with G and can have appreciable greater power than with 4 replicates. The value of $K_o(G)$ is not substantially affected by whether the MOI is dominant (see scenarios 8–10) or recessive (see scenarios 1–7).

### *Regression modeling results*

We use ordinary least squares (OLS) regression analysis with power at the 0.0001 significance level as the dependent variable for each of the $4^4 \times 2^3 \times 3 \times 5$ (30720) model specifications. We consider the 9 factors listed in Table 3 and all possible two-way combinations in our regression model to assess the relative importance of the factors in determination of power to detect association. We also use the square of the number of replicates to model the optimal number of replicates. The analysis finds a significant fit ($F_{55,30664} = 1348.07$, p-value < 0.0001) with $R^2$ equal to 0.71. Genotype relative risk ($R_2$) has the largest *F*-statistic (34333.5 with 1 df), with increasing $R_2$ associated with greater power. Sample size has the second largest *F*-statistic (15002.4 with 1 df). The MOI also has a highly significant *F*-statistic (5869.7 with 2 df). For a fixed genotype relative risk $R_{2,}$ the median power is greatest for dominant MOI, followed by multiplicative and then recessive MOIs. The prevalence of the disease ($\phi$), the minor marker allele frequency, and the measurement quality of the pooling are the factors that have the smallest *F*-statistic values. Measurement error explains less of the variance than genetic parameters. In general, increased measurement error reduces the power of the procedure. Further, with genetic parameters fixed, the decrease in power from increased measurement error can be offset either by an increase in *K* or decrease of the number of individuals in each pool.

Among the interaction terms not involving *K*, $p_d \times q_1$, $p_d \times MOI$, $R_2 \times T$, $p_d \times R_2$, and $N \times T$ are highly significant (sorted in increasing *P*-values). The most significant inter-

**Table 2: Maximum power as a function of the number of genotyping($G = J \times K$), number of replicates giving maximum power ($K_o(G)$), number of replicates ($K$) at 95% of the maximum power at specific experimental and genetic parameters and the power at $K = 1$ when assuming no machine replicability variability ($m = 1$)**

| Situation | *N* | MOI | $R_2$ | $\alpha$ | *m* | *MAF* | $p_r$ | G = 40 | G = 80 | G = 160 | G = 320 | G = 640 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 10000 | R | 2.2 | 0.0001 | 2.25 | 0.20 | 0.9 | 38%, 2, (2), 82% | 54%, 4, (3–4), 85% | 64%,6, (4–7), 87% | 72%, 10, (5–16), 87% | 77%, 13, (6–27), 88% |
| 2 | 10000 | R | 2.0 | 0.001 | 2.25 | 0.20 | 0.9 | 43%, 2, (2), 79% | 56%, 4, (3–4), 81% | 65%, 7, (4–7), 82% | 71%, 11, (6–16), 83% | 75%, 16, (7–32), 83% |
| 3 | 10000 | R | 1.8 | 0.01 | 2.25 | 0.20 | 0.9 | 50%, 2, (2), 79% | 62%, 4, (3–4), 80% | 68%, 7, (5–7), 80% | 72%, 13, (6–16), 80% | 75%, 20, (7–32), 81% |
| 4 | 10000 | R | 2.2 | 0.0001 | 2.25 | 0.15 | 1 | 69%, 2, (2), 98% | 84%, 4, (3–4), 99% | 90%, 6, (3–7), 99% | 95%, 10, (4–16), 99% | 96%, 13, (4–32), 99% |
| 5 | 10000 | R | 2.2 | 0.0001 | 2.0 | 0.15 | 1 | 75%, 2, (2), 98% | 87%, 4, (2–4), 99% | 92%, 5, (3–7), 99% | 95%, 8, (3–16), 99% | 97%, 12 (3–32), 99% |
| 6 | 10000 | R | 2.0 | 0.0001 | 2.25 | 0.15 | 1 | 43%,2, (2), 86% | 59%, 4, (3–4), 89% | 70%, 6, (4–7), 90% | 77%, 10, (5–16), 91% | 82%, 13, (6–28), 91% |
| 7 | 10000 | R | 2.0 | 0.0001 | 2.0 | 0.15 | 1 | 49%, 2 (2), 86% | 63%, 4, (3–4), 89% | 73%, 5, (4–7), 90% | 79%, 8, (4–16), 91% | 83%, 12, (5–27), 91% |
| 8 | 2000 | D | 1.5 | 0.0001 | 2.25 | 0.15 | 0.9 | 57%, 3, (2–3), 94% | 73%, 4, (3–6), 96% | 82%, 6, (4–10), 96% | 88%, 10, (4–18), -- | 91%, 13, (5–32), -- |
| 9 | 2000 | D | 1.5 | 0.0001 | 2.25 | 0.15 | 1 | 65%, 3, (2–3), 97% | 80%, 4, (3–6),98% | 88%, 6, (3–9), 98 | 92%, 10, (4–21), -- | 95%, 13, (4–40), -- |
| 10 | 2000 | D | 1.5 | 0.0001 | 2.0 | 0.15 | 1 | 70%, 2, (2–3), 97% | 83%, 4, (3–6), 98% | 90%, 5. (3–10), 98% | 94%, 8, (3–20), -- | 96%, 12, (4–37), -- |

We only consider designs in which the pool size (*T*) is between 10 and 500. The prevalence $\phi$ is 0.05 and disease allele frequency $p_d$ is 0.15.
--: the size of pool is out of our consideration, no power is provided;
*N*: sample size in cases or controls;
MOI: mode of inheritance (R = recessive MOI and D= dominant MOI);
$R_2$: relative risk for subjects homozygous for disease allele;
$\alpha$: significance level;
*m*: machine replicability variance factor;
*MAF*: minor SNP marker allele frequency;
$p_r$: measure of linkage disequilibrium.

**Table 3: List of parameters considered in the multiple regression analysis**

| Parameter | Description | Value |
|---|---|---|
| $N$ | Number of case (control) subjects | 1000, 2000, 5000, 10000 |
| $\phi$ | Prevalence of the disease | 0.01, 0.1 |
| $T$ | Size of the pool | 25, 50, 100, 250, 500 |
| $K$ | Number of replicates of each pool | 1, 2, 4, 8 |
| $p_d$ | Disease allele frequency | 0.1, 0.25 |
| MOI | Modes of inheritance | dominant, recessive, multiplicative |
| $R_2$ | Genotype relative risk of homozygote of disease allele | *1.2, 1.5, 2.25, 4 |
| $q_1$ | Minor SNP marker allele frequency | 0.1, 0.35 |
| $M$ | Machine replicability variance factor | 2.05, 2.1, 2.25, 3 |

*$R_1$ is obtained according to the relationship between $R_1$ and $R_2$; that is for multiplicative MOI, $R_2 = R_1^2$ ; dominant MOI, $R_1 = R_2$; recessive MOI, $R_1 = 1$. We considered all 30720 ($4^4 \times 2^3 \times 3 \times 5$) situations generated from the parameters listed above.

action term is $p_d \times q_1$. This finding is not surprising as there has been extensive documentation in the statistical genetics literature that power for genetic association is maximized when the difference between the disease allele frequency and the SNP marker allele frequency in coupling with the disease allele is 0, with decreasing power occurring as the difference increases [20-23]. The finding of a significant interaction $p_d \times MOI$ between disease allele frequency and disease MOI has also been documented previously, most recently in the work by Skol et al. [4]. The finding underscores the fact that, when all other factors are fixed, the disease allele frequency that gives optimal power differs depending upon the disease MOI.

## Discussion

Our results have produced two types of conclusions. The first is that the genetic parameters of the disease being studied are the most important determinants of the power to detect association. This fact is consistent with the association of ApoE with late onset Alzheimer's Disease [24] and recent association results for age-related macular degeneration [1,3]. In each of these studies, estimated genotype relative risks are approximately 3 for the heterozygote and greater than 9 for the homozygote. In all studies, highly significant associations were observed with less than 500 total cases and controls. Furthermore, for age-related macular degeneration [24], associations were observed for SNP alleles in linkage disequilibrium (LD) with the functional variants. The results from the OLS regression analysis are consistent with this history. The genetic relative risk is the most significant parameter, followed by the sample size. For a fixed genotype relative risk $R_{2,}$ the median power is greatest for dominant MOI, followed by multiplicative and then recessive MOIs. The linear and quadratic terms in the number of replicates $K$ and a number of interactions with $K$ are significant. Since there is an optimal setting of $K$, this result is expected.

The second type of conclusion is guidance about the choice of the number of genotypings $G = J \times K$ and the simultaneous setting of the number of replicates $K$ of the $J$ pools. We have shown that the number of genotypings $G = J \times K$ should be as large as possible (holding all other factors constant) to have the greatest power. When $G$ is fixed, we have shown that there is a setting $K_o(G)$ that maximizes the power when all genetic model parameters are specified. The optimal setting increases as $G$ increases. These differences are practically important and suggest that those conducting pooled studies use the program available from the corresponding author to determine optimal settings. In all situations studied, for fixed value of $G$, power is relatively insensitive to choice of $K$ near $K_o(G)$. Further, when the machine replicability variance factor $m$ is larger than 1, the setting $K = 1$ has power much less than replicated designs. This suggests that such extensions of these designs as staggered nested designs [18] may have little value in genetic pooling studies.

Our work provides the basis for extending recommendations such as those of Sham et al. [10] to include genetic model parameters. For the very large studies possible with pooling, there is strong evidence that increasing the number of genotypings and increasing the number of replicate measurements of each pool can increase power noticeably. This approach is dependent on the assumption that $E(\Pi_i) = E(Y_{ijk})$, where $\Pi_i$ is the fraction of the major allele 2 in a randomly selected subject from the $i^{\text{th}}$ group; that is, the pooled estimate of the intensity of an allele is in fact an unbiased estimate of the allele 2 frequency. Further work will incorporate designs that formally include validation of this assumption.

## Conclusion

Our work extends that of previous researchers who have considered power and sample size calculations for genetic association studies with pooled DNA samples (e.g., [16]). Our extension involves inclusion of genetic model parameters such as disease MOI, disease allele frequency, disease prevalence, marker allele frequency, and genotype relative risks. It is clear from the results of our regression analysis that incorporation of such parameters is important in the design of more powerful genetic association tests. We recommend that researchers incorporate information into their power and sample size calculations for genetic association with pooled DNA, such as choice of numbers of genotypings and the number of replicates that can increase power from such relatively low levels as 40% to 50% to 75% to 80% using the same cases and controls.

## Methods

### *Definitions*

*N*: number of case (control) subjects; we assume equal numbers of cases and controls (balanced design).

*J*: number of pools; $J \geq 2$.

$T = N/J$: number of individuals in each pool; we assume that case subjects are assigned randomly to case pools and control subjects are assigned randomly to control pools.

*K*: number of replicates of each pool; we assume that there is no reassignment of subjects in the replications.

$G = J \times K$: number of case (control) genotypings.

### *Genetic model parameters*

We consider a disease associated with a di-allelic gene with allele *d* associated with increased risk of disease and allele + associated with no increased risk.

$p_d$: allele frequency of disease locus *d* allele.

$p_+ = 1 - p_d$: allele frequency of disease locus wild-type (+) allele.

$\phi$: prevalence of the disease.

$f_2$: probability of having disease with 2 disease alleles in the genotype = penetrance of *dd*.

$f_1$: probability of having disease with 1 disease allele in the genotype = penetrance of *d*+.

$f_0$: probability of having disease with 0 disease alleles in the genotype = penetrance of ++.

### *Genotype relative risks (GRR)*

$$R_2 = \frac{f_2}{f_0}.$$

$$R_1 = \frac{f_1}{f_0}.$$

### *Modes of Inheritance (MOI)*

The three MOIs are characterized by the parameter *R*.

Multiplicative MOI: The penetrances satisfy the equation $R = \frac{f_1}{f_0} = \frac{f_2}{f_1}$; that is, $R_2 = R_1^2$.

Dominant MOI: $R = R_1 = R_2$.

Recessive MOI: $R = R_2 = \frac{R_2}{R_1}$; that is, $R_1 = 1$.

### *SNP marker parameters*

$q_1$: allele frequency of minor SNP marker allele 1 (that is, $0 < q_1 \leq 0.5$).

$q_2$: the frequency of the major SNP marker allele 2.

### *Disequilibrium parameters*

$D_{\max} = \min(p_d q_2, p_+ q_1)$.

$$R_{\max}^2 = \frac{D_{\max}^2}{p_d p_+ q_1 q_2} \text{ (see, e.g., [25]).}$$

$p_r$: measure of linkage disequilibrium between disease gene and SNP marker; here it is a fraction of $R_{\max}^2$ ($0 < p_r \leq 1$); the examples use $p_r = 0.9$.

The detailed computation of case and control genotype probabilities which are functions of the disease allele frequency, minor SNP allele frequency, and linkage disequilibrium parameters are documented in the PAWE-3D Helpfile [19].

We use method [26] implemented in the PAWE software [27] to calculate the power of the 2 × 2 test of independence when each subject is individually genotyped and we report these value in Figures 1 and 2.

### *Case-control frequency of allele 2*

$\Pi_i$: the fraction of the major allele 2 in a randomly selected subject from the $i^{\text{th}}$ group, $i = 0$ for cases, $i = 1$ for controls. It follows that the expectation of $\Pi_i$ is given by:

$$E(\Pi_i) = \frac{1}{2} P_{i1} + P_{i2},$$

where $P_{i1}$ is the frequency of the heterozygous genotype with allele 2 in the $i^{\text{th}}$ group and $P_{i2}$ is the frequency of the homozygous genotype with allele 2 in the $i^{\text{th}}$ group. In addition,

$$\mathrm{var}(\Pi_i) = \tau_i^2 = E(\Pi_i^2) - [E(\Pi_i)]^2 = \frac{1}{4}P_{i1}(1-P_{i1}) + P_{i2}(1-P_{i2}) - P_{i1}P_{i2}.$$

### *Analysis of variance (ANOVA) table for two-stage nested design*

*Specification of ANOVA model*

$A_{ijk}$: intensity level of allele 2 in the $i^{\text{th}}$ group ($i$ = 0 for cases, 1 for controls), $j^{\text{th}}$ pool ($j$ = 1,...,$J$), $k^{\text{th}}$ replicate ($k$ = 1,...,$K$).

$B_{ijk}$: intensity level of allele 1 in the $i^{\text{th}}$ group ($i$ = 0 for cases, 1 for controls), $j^{\text{th}}$ pool ($j$ = 1,..., $J$), $k^{\text{th}}$ replicate ($k$ = 1,..., $K$).

$Y_{ijk} = \dfrac{A_{ijk}}{A_{ijk} + B_{ijk}}$ : fraction of SNP allele 2 estimated in the $i^{\text{th}}$ group (i = 0 for cases, 1 for controls), $j^{\text{th}}$ pool ($j$ = 1,..., $J$), $k^{\text{th}}$ replicate ($k$ = 1,..., $K$).

Model:

$$Y_{ijk} = \mu + \alpha_i + P_{j(i)} + \sigma_E E_{ijk},$$

where the case or control effect is $\alpha_i = E(\Pi_i) - E(\dfrac{\Pi_0 + \Pi_1}{2})$, $i$ = 0,1, subject to the constraint $\sum\alpha_i = 0$. The random sampling effect of the allele 2 frequency associated with the $j^{\text{th}}$ pool in either cases or controls is $P_{j(i)} \sim N(0, \sigma_{P,i}^2)$, with $\sigma_{P,i}^2 = \dfrac{\tau_i^2}{T} = \dfrac{J\tau_i^2}{N}$. Finally, $\{E_{ijk}\}$ are independent $N(0,1)$ random variables incorporating the additional variability due to the measurement process. See below for more details regarding the specification $\sigma_{P,i}^2 = \dfrac{\tau_i^2}{T}$. It follows that

$$\mathrm{var}(Y_{ijk}) = \sigma_{P,i}^2 + \sigma_E^2 = \frac{\tau_i^2}{T} + \sigma_E^2 = \frac{J\tau_i^2}{N} + \sigma_E^2.$$

Here, var($Y_{ijk}$) is modeled as the sum of two components of variance. The first, $\sigma_{P,i}^2 = \dfrac{J\tau_i^2}{N}$, is due to the sampling variation of the frequency of allele 2 in the subjects assigned to each pool. The second, $\sigma_E^2$, is due to the measurement error of the processing of the pooled material. Under an ideal measurement process, $\sigma_E^2$ = 0; we define a parameter $m$ to capture the departure from this ideal. The parameter $m$ (machine replicability variance factor), $m \geq 1$, is defined by $\sigma_E^2 = \sigma_{P,i}^2(m-1)$, so that $m$ = 1 represents the ideal measurement process and $m$ > 1 models additional variability due to a less than perfect measurement process. The fraction of var($Y_{ijk}$) due to the measurement process is $\dfrac{m-1}{m}$.

This model is dependent on the assumption that $E(\Pi_i) = E(\dfrac{A_{ijk}}{A_{ijk} + B_{ijk}}) = E(Y_{ijk})$. Also, let $\rho = \dfrac{\max(\sigma_{P,0}^2, \sigma_{P,1}^2)}{\min(\sigma_{P,0}^2, \sigma_{P,1}^2)}$. This value is an indication of the adequacy of the approximation of the NCP in equation (1) below [28].

Let

$$Y_{ij\bullet} = \frac{\sum_{k=1}^{K} Y_{ijk}}{K}, Y_{i\bullet\bullet} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}}{JK}, \text{ and } Y_{\bullet\bullet\bullet} = \frac{\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}}{2JK}.$$

Following Scheffé [29], the means used in the sums of squares can be expressed in terms of the ANOVA model as

$$Y_{ij\bullet} = \frac{\sum_{k=1}^{K} Y_{ijk}}{K} = \mu + \alpha_i + P_{j(i)} + \sigma_E E_{ij\bullet}, \text{ where } E_{ij\bullet} = \sum_{k=1}^{K} E_{ijk} / K;$$

$$Y_{i\bullet\bullet} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}}{JK} = \mu + \alpha_i + P_{\bullet(i)} + \sigma_E E_{i\bullet\bullet}, \text{ where } P_{\bullet(i)} = \sum_{j=1}^{J} P_{j(i)} / J \text{ and } E_{i\bullet\bullet} = \sum_{j=1}^{J}\sum_{k=1}^{K} E_{ijk} / JK;$$

$$Y_{\bullet\bullet\bullet} = \frac{\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}}{2JK} = \mu + 0 + P_{\bullet(\bullet)} + \sigma_E E_{\bullet\bullet\bullet}, \text{ where } P_{\bullet(\bullet)} = \sum_{i=0}^{1} P_{\bullet(i)} / 2 \text{ and } E_{\bullet\bullet\bullet} = \sum_{i=0}^{1} E_{i\bullet\bullet} / 2.$$

Then,

$$Y_{ij\bullet} - Y_{i\bullet\bullet} = P_{j(i)} + \sigma_E E_{ij\bullet} - (P_{\bullet(i)} + \sigma_E E_{i\bullet\bullet}) = T_{ij} - T_{i\bullet},$$

where

$$T_{ij} = P_{j(i)} + \sigma_E E_{ij\bullet}, T_{i\bullet} = \sum_{j=1}^{J} T_{ij} / J, \text{ and } T_{ij} \sim N(0, \sigma_{P,i}^2 + \frac{\sigma_E^2}{K})$$

.

If we let $W_i$ represent $n$ independent and identically distributed $N(\mu, \sigma^2)$ random variables, then $\sum_{i=1}^{n}(W_i - \bar{W})^2$ has the distribution $\sigma^2 \chi^2_{n-1}$ [18]. Consequently,

$$
\begin{aligned}
SS_P &= K\sum_i \sum_j (Y_{ij\bullet} - Y_{i\bullet\bullet})^2 \\
&= K\sum_i [\sum_j (T_{ij} - T_{i\bullet})^2] \\
&= K\sum_i (\sigma^2_{P,i} + \frac{\sigma^2_E}{K})X_i,
\end{aligned}
$$

where $X_i \sim \chi^2_{J-1}$. The sum of squares $SS_P$ therefore has the distribution $K(\sigma^2_{\bar{P}} + \frac{\sigma^2_E}{K})\chi^2_{2(J-1)}$ when the null hypothesis is true, with $\sigma^2_{\bar{P}} = \sigma^2_{P,0} = \sigma^2_{P,1}$. Further $E(SS_P) = 2(J-1)(K\sigma^2_{\bar{P}} + \sigma^2_E)$ under both the null and alternative hypotheses with $\sigma^2_{\bar{P}} = \frac{J\bar{\tau}^2}{N}$ and $\bar{\tau}^2 = \frac{\sum_{i=0}^{1}\tau_i^2}{2} = \frac{\sum \mathrm{var}(\Pi_i)}{2}$. The distribution of $SS_P$ under the alternative is a weighted sum of independent central chi-square distributions.

To obtain the distribution of $SS_A$, consider

$$
Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet} = (\alpha_i + P_{\bullet(i)} + \sigma_E E_{i\bullet\bullet}) - (P_{\bullet(\bullet)} + \sigma_E E_{\bullet\bullet\bullet}) = \alpha_i + S_i - S_\bullet,
$$

where $S_i = P_{\bullet(i)} + \sigma_E E_{i\bullet\bullet}$ with $S_i \sim N(0, \frac{\sigma^2_{P,i}}{J} + \frac{\sigma^2_E}{JK})$. Then,

$$
SS_A = JK\sum_i (Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet})^2 = JK\sum_i (\alpha_i + S_i - S_\bullet)^2.
$$

The null distribution of $SS_A$ is a scaled central chi-squared random variable with scaling factor $\sigma^2_E + K\sigma^2_{\bar{P}}$ so that

$$
F_{1,2(J-1)} = \frac{SS_A/1}{SS_P/2(J-1)}
$$

has a central $F$-distribution with 1 numerator degree of freedom and $2(J - 1)$ denominator df when $H_0$: $\alpha_i \equiv 0$ is valid. Under the alternative hypothesis, the distribution of $SS_A$ is a weighted sum of non-central chi-squared random variables. The approximation to the alternative distribution of the $F$-test proposed here is that it is a non-central $F$ with 1 numerator degree of freedom, $2(J - 1)$ denominator df, and non-centrality parameter (NCP) $\delta^2$, where

$$
\delta^2 = \frac{JK\sum \alpha_i^2}{(K\sigma^2_{\bar{P}} + \sigma^2_E)} = \frac{NK\sum \alpha_i^2}{T(K\sigma^2_{\bar{P}} + \sigma^2_E)}. \qquad (1)
$$

As shown by Gronow [28], the inequality in variance does not affect the power approximation when $p \leq 1.5$ Since $\sigma^2_E = \frac{\bar{\tau}^2}{T}(m-1)$, where $1 \leq m$,

$$
\delta^2 = \frac{JK\sum \alpha_i^2}{(K\sigma^2_{\bar{P}} + \sigma^2_E)} = \frac{NK\sum \alpha_i^2}{T[K\frac{\bar{\tau}^2}{T} + \frac{\bar{\tau}^2}{T}(m-1)]} = \frac{NK\sum \alpha_i^2}{[K\bar{\tau}^2 + \bar{\tau}^2(m-1)]},
$$

which is not dependent upon $J$, assuming this model. This result is due to the fact that we assumed $\sigma^2_{P,i} = \frac{\tau_i^2}{T}$, which is an assumption that each individual's variance contributes equally to the variance of the pool. The factor $(m - 1)$ includes the cumulative effect of such sources of additional variability as experimental error, differential variability in processing of individuals, and other sources.

### *Multiple regression analysis of approximate power*

We calculated the approximate power of the experimental design under various values of parameters (Table 3). We then used OLS multiple regression analysis to identify the parameters that had the greatest impact on power, using SAS software [30]. For independent variables, we used all variables listed in Table 3, all two way interactions of these variables, and $K^2$, the square of the number of replicates to incorporate the existence of an optimal number of replicates. We considered type I errors at 0.01, 0.001 and 0.0001 levels. It might be argued that researchers should use 0.0001 or less as a stringent significance level if the design is applied in a genome-wide association study. Since DNA pooling techniques are normally used as 1st stage screening and for 1st stage design, researchers may be more concerned with false negatives than false positives [9,31].

## Authors' contributions

FJ, SJF, NRM, and DG conceived of the study design. FJ performed all statistical analyses presented in the manuscript. CH wrote software to assist FJ in her analyses. FJ, SJF, and DG wrote the original manuscript and all revisions. All authors have read and approve the final manuscript.

## Acknowledgements

This research was supported by the US National Institutes of Health grants NIMH R01 MH071523 (NRM), NIMH 2R01 MH04480114A1 (SJF), and MH44292 (FJ). The authors thank Dr. Yaning Yang for providing comments on our revised version of the manuscript.

## References

1. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, Sangiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J: **Complement factor H polymorphism in age-related macular degeneration.** *Science* 2005, **308(5720):**385-389.
2. Ozaki K, Tanaka T: **Genome-wide association study to identify SNPs conferring risk of myocardial infarction and their functional analyses.** *Cell Mol Life Sci* 2005, **62(16):**1804-1813.
3. Dewan A, Liu M, Hartman S, Zhang SS, Liu DT, Zhao C, Tam PO, Chan WM, Lam DS, Snyder M, Barnstable C, Pang CP, Hoh J: **HTRA1 promoter polymorphism in wet age-related macular degeneration.** *Science* 2006, **314(5801):**989-992.
4. Skol AD, Scott LJ, Abecasis GR, Boehnke M: **Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies.** *Nat Genet* 2006, **38(2):**209-213.
5. Wang H, Thomas DC, Pe'er I, Stram DO: **Optimal two-stage genotyping designs for genome-wide association scans.** *Genet Epidemiol* 2006, **30(4):**356-368.
6. Schaid DJ, Sommer SS: **Genotype relative risks: methods for design and analysis of candidate-gene association studies.** *Am J Hum Genet* 1993, **53(5):**1114-1126.
7. Purcell S, Cherny SS, Sham PC: **Genetic power calculator: design of linkage and association genetic mapping studies of complex traits.** *Bioinformatics* 2003, **19(1):**149-150.
8. Gordon D, Haynes C, Blumenfeld J, Finch SJ: **PAWE-3D: visualizing power for association with error in case-control genetic studies of complex traits.** *Bioinformatics* 2005, **21(20):**3935-3937.
9. Satagopan JM, Elston RC: **Optimal two-stage genotyping in population-based association studies.** *Genet Epidemiol* 2003, **25(2):**149-157.
10. Sham P, Bader JS, Craig I, O'Donovan M, Owen M: **DNA Pooling: a tool for large-scale association studies.** *Nat Rev Genet* 2002, **3(11):**862-871.
11. Kirov G, Nikolov I, Georgieva L, Moskvina V, Owen MJ, O'Donovan MC: **Pooled DNA genotyping on Affymetrix SNP genotyping arrays.** *BMC Genomics* 2006, **7:**27.
12. Meaburn E, Butcher LM, Liu L, Fernandes C, Hansen V, Al-Chalabi A, Plomin R, Craig I, Schalkwyk LC: **Genotyping DNA pools on microarrays: tackling the QTL problem of large samples and large numbers of SNPs.** *BMC Genomics* 2005, **6(1):**52.
13. Allison DB, Schork NJ: **Selected methodological issues in meiotic mapping of obesity genes in humans: issues of power and efficiency.** *Behav Genet* 1997, **27(4):**401-421.
14. Simonic I, Gericke GS, Ott J, Weber JL: **Identification of genetic markers associated with Gilles de la Tourette syndrome in an Afrikaner population.** *Am J Hum Genet* 1998, **63(3):**839-846.
15. Simonic I, Nyholt DR, Gericke GS, Gordon D, Matsumoto N, Ledbetter DH, Ott J, Weber JL: **Further evidence for linkage of Gilles de la Tourette syndrome (GTS) susceptibility loci on chromosomes 2p11, 8q22 and 11q23-24 in South African Afrikaners.** *Am J Med Genet* 2001, **105(2):**163-167.
16. Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG: **Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design.** *Ann Hum Genet* 2002, **66(Pt 5-6):**393-405.
17. Le Hellard S, Ballereau SJ, Visscher PM, Torrance HS, Pinson J, Morris SW, Thomson ML, Semple CA, Muir WJ, Blackwood DH, Porteous DJ, Evans KL: **SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis.** *Nucleic Acids Res* 2002, **30(15):**e74.
18. Montgomery DC: **Design and Analysis of Experiments.** Sixth edition. Hoboken , J. Wiley and Sons; 2005.
19. **PAWE-3D** [http://linkage.rockefeller.edu/pawe3d/]
20. Zondervan KT, Cardon LR: **The complex interplay among factors that influence allelic association.** *Nat Rev Genet* 2004, **5(2):**89-100.
21. Pfeiffer RM, Gail MH: **Sample size calculations for population- and family-based case-control association studies on marker genotypes.** *Genet Epidemiol* 2003, **25(2):**136-148.
22. Ji F, Yang Y, Haynes C, Finch SJ, Gordon D: **Computing asymptotic power and sample size for case-control genetic association studies in the presence of phenotype and/or genotype misclassification errors.** *Stat Appl Genet Mol Biol* 2005, **4(1):**Article 37.
23. Gordon D, Finch SJ: **Factors affecting statistical power in the detection of genetic association.** *J Clin Invest* 2005, **115:**1408-1418.
24. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA: **Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families.** *Science* 1993, **261(5123):**921-923.
25. Pritchard JK, Przeworski M: **Linkage disequilibrium in humans: models and data.** *Am J Hum Genet* 2001, **69(1):**1-14.
26. Gordon D, Finch SJ, Nothnagel M, Ott J: **Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms.** *Hum Hered* 2002, **54(1):**22-33.
27. **PAWE** [http://linkage.rockefeller.edu/pawe/]
28. Gronow DG: **Test for the significance of the difference between means in two normal populations having unequal variances.** *Biometrika* 1951, **38(1-2):**252-256.
29. Scheffe H: **The Analysis of Variance.** In *Wiley Classics Library* New York , Wiley-Interscience; 1999:477.
30. **SAS, version 9.1** [http://www.sas.com]
31. Elston RC, Guo X, Williams LV: **Two-stage global search designs for linkage analysis using pairs of affected relatives.** *Genet Epidemiol* 1996, **13(6):**535-558.