Published in final edited form as:

Ann Hum Genet. 2014 September ; 78(5): 345–356. doi:10.1111/ahg.12073.

Disentangling Pooled Triad Genotypes for Association Studies

Min Shi¹, David M. Umbach¹, and Clarice R. Weinberg¹

¹Biostatistics Branch, NIEHS, NIH, DHHS, Research Triangle Park NC

Abstract

Association studies that genotype affected offspring and their parents (triads) offer robustness to genetic population structure while enabling assessments of maternal effects, parent-of-origin effects, and gene-by-environment interaction. We propose case-parents designs that use pooled DNA specimens to make economical use of limited available specimens. One can markedly reduce the number of genotyping assays required by randomly partitioning the case-parent triads into pooling sets of h triads each and creating three pools from every pooling set, one pool each for mothers, fathers, and offspring. Maximum-likelihood estimation of relative risk parameters proceeds via log-linear modeling using the expectation-maximization algorithm. The approach can assess offspring and maternal genetic effects and accommodate genotyping errors and missing genotypes. We compare the power of our proposed analysis for testing offspring and maternal genetic effects to that based on a difference approach considered by Lee and that of the gold-standard based on individual genotypes, under a range of allele frequencies, missing parent proportions and genotyping error rates. Power calculations show that the pooling strategies cause only modest reductions in power if genotyping errors are low, while reducing genotyping costs and conserving limited specimens.

Introduction

In searching for genes related to young-onset diseases, genetic epidemiologists often employ case-parents designs, which call for genotyping affected offspring and their biological parents (triads) (Weinberg *et al.*, 1998, Spielman *et al.*, 1993). The case-parents approach may be preferred to the case-control approach to ensure that inferences about genetic associations are robust to hidden genetic population structure and to enable exploration of maternal genetic effects (Wilcox *et al.*, 1998), fetal-maternal interactions (Sinsheimer *et al.*, 2003) and parent-of-origin effects (Weinberg, 1999). Robustness for assessing effects of inherited alleles arises because those inferences are conditional on parental genotypes, but that robustness comes at a cost. A case-parents design needs to genotype two controls (the biological parents) for each affected individual rather than the one control typical with case-control designs. To reduce the genotyping needed for family-based association studies while preserving robustness and statistical efficiency, several authors proposed comparing variant allele frequencies measured in pooled DNA specimens (Zou & Zhao, 2005, Risch & Teng, 1998, Bader & Sham, 2002, Lee, 2005, Beckman *et al.*, 2006). Here we propose a method to analyze pooled DNA specimens based on probabilistically disaggregating the pooled genotypes into their component individual genotypes. We assume that the locus is diallelic, the site of a single nucleotide polymorphism or SNP, and that the assay can effectively count the number of copies of the variant autosomal allele in pooled DNA from a few individuals.

We consider the following pooling strategy. First, randomly partition the triads into pooling sets of size h . For each set, pool equal amounts of DNA from each mother and genotype the mother pool with an assay that reports the number of variant alleles as an integer from 0 to $2h$. Do likewise for the h fathers and the h children. This strategy reduces the genotypes required for each pooling set from $3h$ to three.

Using a similar strategy, Lee (2005) proposed a test of offspring genetic effects based on comparing allele frequencies measured in the pools without any attempt to computationally disaggregate the pooled triad genotypes into individual triad genotypes. His test is simple to carry out and maintains the nominal Type I error rate even for error-prone genotyping assays and in a stratified population (Lee, 2005). It, however, sacrifices other features provided by log-linear models, such as relative risk estimation, and tests of maternal or parent-of-origin effects. Our alternative approach is based on probabilistically disaggregating the pooled triad genotypes into individual triad genotypes when fitting log-linear models via the expectation maximization (EM) algorithm (Dempster *et al.*, 1977). Though more computationally intensive than Lee's, our approach enables relative risk estimation and allows for any tests of specific genetic mechanisms that a log-linear model can provide. Unlike Lee's, our approach maintains the correct Type I error rate in the presence of missing genotypes.

Methods

Disaggregation without genotyping errors

Assume that we have ascertained N case-parent triads and are using the pooling strategy described above. Thus, we form a total of $n = N/h$ sets of three DNA pools, each set consisting of a mother, a father and an offspring pool of h individuals each. (N need not be a multiple of h , as pooling sets can differ in size, but for simplicity of exposition and programming we make that assumption here and in the software that we make available.)

We index subjects according to their pooling set $k \in \{1, 2, 3, \dots, n\}$, their family within pooling set $j \in \{1, 2, 3, \dots, h\}$, and their family relationship $i \in \{m, f, c\}$ (corresponding to *mother*, *father* or *child*). Accordingly, we represent the genotype of a triad at a single SNP by the 1×3 vector $\mathbf{G}_{kj} = (g_{kjm}, g_{kjf}, g_{kjc})$, where the g_{kji} denote the number of copies of the variant allele, *i.e.*, 0, 1 or 2, at that locus for each family member. Further, we call the unobserved $1 \times 3h$ vector of the triad genotypes for all h families in pooling set k a triad configuration, and denote it as $\mathbf{C}_k = (\mathbf{G}_{k1}, \mathbf{G}_{k2}, \mathbf{G}_{k3}, \dots, \mathbf{G}_{kh})$. We assume that for each diallelic locus the assay correctly reports the number of variant alleles in a pool so that the genotype reported for a pool of size h is the sum of the genotypes of the h individuals in the

pool and ranges from 0 to $2h$. Let $T_{ki} = \sum_{j=1}^h g_{jki}$ be the sum of genotypes for family member i in pooling set k , and let $\mathbf{T}_k = (T_{km}, T_{kf}, T_{kc})$ denote the 1×3 vector containing those genotype sums. We refer to $\mathbf{T}_k$ as the pooling set genotype. If we had observed the individual triad genotypes, we could fit log-linear models directly. Instead, we observe the incomplete data $\mathbf{T}_k$ and employ the EM algorithm by analyzing the unobserved $\mathbf{G}_{kj}$ as pseudo-complete data. Initially we assume that every $\mathbf{T}_k$ is measured without error and thus

provides the actual sums of the genotypes in pooling set k . Later we will introduce genotyping errors and will distinguish between observed and true values of T_k .

A correctly observed value of T_k can arise from one or more different triad configurations. Let $C = (G_1, G_2, G_3, \dots, G_h)$ denote a hypothetical vector of genotypes for h triads. We regard triad configuration C as *compatible* with pooling set genotype T_k if the mother, father, and child genotypes, respectively, in C sum to the corresponding entries in T_k . Because the ordering in C is arbitrary, permutation of the G_j in a compatible configuration yields another compatible configuration. For example, for $h = 2$, the triad configurations $C = ((2,2,2), (1,2,2))$ and $C = ((1,2,2), (2,2,2))$ both yield $T_k = (3,4,4)$. More importantly, distinct lists of G_j can be compatible. For example, $C = ((2,2,2), (0,0,0))$ and $C = ((1,1,1), (1,1,1))$ both yield $T_k = (2,2,2)$. For each T_k one can identify the set of all compatible triad configurations; denote that set by $\mathbb{C}(T_k)$, to emphasize that the list of possible compatible triad configurations depends on the observed pooling set genotype vector. Every triad configuration in each $\mathbb{C}(T_k)$ is assigned an initial weight (e.g., equal weights), normalized so that the weights for the configurations in each $\mathbb{C}(T_k)$ sum to 1. These weights, essentially estimates of the probability that C is the true triad configuration among those compatible with T_k , are updated iteratively through the EM steps.

The complete data likelihood would be based on the unobserved triad genotypes, G_{kj} , and can be written as follows:

$$\prod_{k=1}^n \prod_{j=1}^h \Pr(G_{kj} | \theta) = \prod_{k=1}^n \prod_{C \in \mathbb{C}(T_k)} \prod_{j=1}^h \Pr(G_{kj} | \theta)^{I(C=C_k^*)}$$

where θ is a parameter vector, and C_k^* is the true triad configuration among those in $\mathbb{C}(T_k)$. The risk model $\Pr(G_{kj} | \theta)$ models the distribution of triad genotypes conditional on the offspring's being affected. We use a robust log-linear risk model where θ includes mating-type and relative-risk parameters (Weinberg *et al.*, 1998). With dependence on k and j suppressed for notational simplicity, a version of this model that includes both offspring and maternal genetic effects is:

$$\ln(\Pr(G | \theta)) = \sum_{a=0}^2 \sum_{b=a}^2 \mu_{ab} I([(g_m, g_f) = (a, b)] \text{ or } [(g_m, g_f) = (b, a)]) + \sum_{b=1}^2 \beta_b I(g_c = b) + \sum_{b=1}^2 \alpha_b I(g_m = b) + \ln(2) I_{(G=(1,1,1))}.$$

Here, the μ_{ab} for $a, b \in \{0,1,2\}$ with $b \neq a$ represent 6 mating-type parameters; $e^{ab} \equiv R_b$ for $b \in \{1,2\}$, represent the relative risks for a child carrying b copies of the variant compared to no copies; $e^{ab} \equiv S_b$ for $b \in \{1,2\}$ represent the relative risks for a child whose mother carries b copies of the variant compared to no copies; and $I_{(A)}$ is an indicator function with value 1 if statement A is true and 0 otherwise. Use of 6 mating-type parameters instead of 9 implicitly imposes an assumption of mating symmetry, namely, that in the population $\Pr((g_m, g_f) = (a, b)) = \Pr((g_m, g_f) = (b, a))$ for all $a, b \in \{0,1,2\}$. This codominant risk model can be modified to accommodate dominant, recessive, or log-additive genetic effects, gene-environment interactions, fetal-maternal interactions and parent-of-origin effects.

With pooling, multiple triad configurations may be compatible with an observed T_k . However, the EM algorithm maximizes the likelihood iteratively by estimating the expected log-likelihood. In the expectation step, the indicator function $I(C=C_k^*)$ is replaced by an estimate of its expectation, that is, an estimate of $\Pr(C=C_k^*)$ across the triad configurations compatible with T_k (these probabilities are the weights mentioned earlier). The maximization step then maximizes the resulting pseudo-complete log-likelihood (details in the online Supplement S1).

Briefly, for each observed T_k , we identify the corresponding $\mathbb{C}(T_k)$. The EM algorithm proceeds by fitting the pseudo-complete multinomial data using a log-linear model (Weinberg *et al.*, 1998) for $\Pr(G_{kj} | \theta)$, treating individual triads as independent. At each EM iteration, for each pooling set, all G_{kj} in a given C are assigned that C 's weight; however, a particular value of G_{kj} may occur in multiple C within one $\mathbb{C}(T_k)$. Consequently, each distinct value of G_{kj} is assigned a weight equal to the sum of the weights for all the C in which it occurs for the pooling set. These weights are further summed across pooling sets to generate a pseudo-count (not necessarily an integer) for each possible triad genotype. The maximization step fits the log-linear model to these pseudo-counts to update estimates of both θ (the mating-type and relative-risk parameters) and the probability of each possible G_{kj} for each triad. Imposing independence, the products of the triad probabilities across component compatible triads update estimates of $\Pr(C=C_k^*)$. Following convergence of the EM, the observed data likelihood is maximized using the final estimated parameters $\hat{\theta}$ (Supplement S1):

$$\prod_{k=1}^n \sum_{C \in \mathbb{C}(T_k)} \prod_{j=1}^h \Pr(G_{kj} | \hat{\theta})$$

Disaggregation with genotyping errors

Until now we have considered an idealized scenario with no genotyping errors. More realistically, genotype calls in pooled samples will be subject to error, although we assume individuals are genotyped without error. With genotyping errors present, a measured pooling-set genotype does not necessarily equal the sum of the true genotypes. Allowing for errors greatly expands the number of compatible configurations in $\mathbb{C}(T_k)$. To extend our method to handle genotyping errors, we continue to denote the observed pooling set genotype by $T_k = (T_{km}, T_{kf}, T_{kc})$; but, to account for possible errors, we need to consider a true pooling set genotype denoted $T_k^* = (T_{km}^*, T_{kf}^*, T_{kc}^*)$ and a $(2h+1)$ -by- $(2h+1)$ genotyping error matrix E where each entry $E_{uv} = \Pr(T_{ki} = u | T_{ki}^* = v)$ for $u, v \in \{0, 1, \dots, 2I\}$ and $i \in \{m, f, c\}$. We apply the same genotyping error matrix E independently for pools from any family member. When there are no genotyping errors $T_k = T_k^*$ and E is the identity matrix. When there are genotyping errors, T_k^* is unknown, but we can construct a set, $\mathbb{T}(E, T_k)$, containing all possible candidate “true” pooling set genotypes T_k^E , *i.e.*, those from which the observed T_k could have originated under error model E . Then, $\mathbb{C}(T_k)$ is the union of the $\mathbb{C}(T_k^E)$ over $\mathbb{T}(E, T_k)$. For simplicity, we restrict our error model so that an observed genotype is either the same as the true genotype or differs from it by 1. To model such errors parsimoniously, we

use just two unknown error parameters in $\mathbf{E}$, e_1 and e_2 . Here e_1 is the probability of being off by 1 when the true T_{ki}^* is at either extreme, *i.e.*, 0 or $2h$, and $2e_2$ is the corresponding probability when the true T_{ki}^* is between those extremes, split equally between errors of 1 or -1 (Table 1, $\mathbf{E}$ for $h=2$). An alternative error model allows upward errors in genotype calling (e.g., calling a pooled genotype of 3 as 4) to differ from downward errors (e.g., calling a pooled genotype of 3 as 2). In that scenario, the two unknown error parameters, e_u and e_d , denote the probability of upward and downward error, respectively.

Fitting via the EM algorithm starts with constructing $\mathbb{C}(\mathbf{T}_k)$ by listing $\mathbb{T}(\mathbf{E}, \mathbf{T}_k)$ and every $\mathbb{C}(\mathbf{T}_k^E)$ for each observed pooling-set genotype $\mathbf{T}_k$. We illustrate the construction of $\mathbb{C}(\mathbf{T}_k)$ under our error model for a pool size of 2 when $\mathbf{T}_k = (4, 4, 4)$ (Figure 1). We assume independent genotyping errors in the mother, father and child pools, *i.e.*,

$P(\mathbf{T}_k | \mathbf{T}_k^E = \mathbf{T}_k^*) = \prod_{i \in \{m, f, c\}} P(T_{ki} | T_{ki}^E = T_{ki}^*)$. Although not explicit in our notation, these probabilities depend on the error parameters through our error model, so the EM algorithm updates estimates of error parameters as well as of θ . Following convergence, the observed data likelihood is maximized using the final estimated parameters (Appendix):

$$\prod_{k=1}^n \sum_{\mathbf{T}_k^E \in \mathbb{T}(\mathbf{E}, \mathbf{T}_k)} \sum_{\mathbf{C} \in \mathbb{C}(\mathbf{T}_k^E)} \left(\Pr(\mathbf{T}_k | \mathbf{C} = \mathbf{C}_k^*) \prod_{j=1}^h \Pr(\mathbf{G}_{kj} | \hat{\theta}) \right).$$

R code for performing the EM steps is available upon request from the authors.

Missing Parental Genotypes

Our framework accommodates families with missing parents provided that missingness is noninformative conditional on the observed data. Equivalently, conditional on the genotypes of the offspring and the available parent in a triad, the genotype distribution of missing parents is probabilistically not different from that of observed parents with the same genotypes for spouse and offspring. The random partition of triads into pooling sets needs to be done separately within strata defined by the triads' missing data patterns: no missing parents, missing father, or missing mother. Accordingly, any family missing a parent will only contribute to two pools, instead of to the three pools for complete triads. The only complication is an increased number of compatible triad genotype configurations for families contributing to only two pools. (Although our methodology can, in principle, accommodate families with both parental genotypes missing, such individuals contribute little information and we avoid them in this paper.) The rest of the EM steps proceed as described above.

Modified Lee's Method

Lee (2005) proposed a method applicable to a variety of family structures for testing offspring genetic effects using pooled specimens. His method is based on the difference between the count of transmitted variant alleles and the count of non-transmitted variant alleles or, equivalently, between twice the number of variant alleles carried by the affected offspring and the total number carried by the parents. We restrict consideration to case-

parent triads, using a minor modification of Lee's test statistic that should produce better power (see Supplement S2), and we extend it to allow assessment of maternal genetic effects.

Noncentrality Parameter Calculations

To evaluate power, we computed the noncentrality parameter (NCP) of a chi-square test statistic based on expected counts. The NCP for the likelihood ratio test (Agresti, 1990) is the change in deviance (twice the maximized log likelihood) between models fitted with and without the parameter(s) of interest to "data" that are expected counts calculated under the specified scenario. The NCP for the difference-based method is the square of the Z-statistic calculated from the expected counts. Power is calculated from the NCP based on tail probabilities for a noncentral chi-squared distribution. Because the difference-based approach provides one-degree-of-freedom (df) tests, to make NCPs easy to compare we fit log-linear models specifying log-additive risks so that the likelihood ratio tests also had one df. Even when a specified pooling scenario did not have genotyping errors, we fit a likelihood model that included genotyping-error parameters. We used simulations to confirm certain NCP-based power calculations and to investigate estimation (see Supplement S4 for methods).

The key to NCP calculations is calculation of expected counts based on a specified scenario that fixes the risk model $\Pr(G_{kj} | \theta)$ and its parameters as well as the genotype distribution in the population. Using that information, we computed the probability of each of the 15 possible case-parent-triads. To calculate the expected count of a triad genotype, we multiplied its probability by the number of families under study. The process for pooling set genotypes was similar but more complicated. We first created a list of all possible genotypes for pooling sets of size h by listing all possible sets consisting of h (not necessarily distinct) of the 15 possible case-parent-triad genotypes and summing the component genotypes in each set. Because pooling sets are formed at random, the probability of a given pooling set genotype appearing in a study is the product of the probabilities of the component triad genotypes. The expected count for any pooling set genotype is the product of its probability and the total number of pooling sets. With genotyping errors present, the expected counts just described apply to the "true" pooling set genotypes. We used the specified error matrix to determine what fraction of the expected counts from a "true" pooling set genotype were transferred to each different "observed" pooling set genotype, accumulating the expected counts appropriately from the possible "true" pooling set genotypes.

We considered a study of $N=1000$ triads and a pool size of $h=2$, yielding 500 pooling sets. Usually, we assumed a homogenous population where the SNP under study was in Hardy-Weinberg Equilibrium in the population. (This assumption is convenient for calculations but not needed for validity; we consider stratified populations in Supplement S5.) The risk allele frequency ranged from 0.1 to 0.9 in increments of 0.1. Our populations always exhibited mating symmetry, which is needed only for assessing maternal genetic effects. We assumed a log-linear model for risk and considered scenarios with either offspring genetic effects or maternal genetic effects, but not both simultaneously. The relative-risk parameters ((R_1, R_2) for offspring effects and (S_1, S_2) for maternal effects) were set at (1.2, 1.44). (We consider

other modes of inheritance in Supplements S4 and S5.) We considered scenarios without genotyping errors and scenarios with moderate ($e_1=0.05$, $e_2=0.048$) or high ($e_1=0.1$, $e_2=0.091$) genotyping errors when the error matrix was parameterized using e_1 and e_2 . (We considered scenarios where the error matrix was parameterized using e_u and e_d , ($e_u=0$, $e_d=0$), ($e_u=0.05$, $e_d=0.05$), ($e_u=0.1$, $e_d=0.05$) in Supplement S3 and a few more complicated error scenarios in Supplement S7.) We also considered scenarios where fathers in 50% of the families were unavailable.

NCPs permit easy calculation of power for any other sample size. For example, to calculate power for a different number of pools of size $h=2$, say M , instead of the 500 that we used, one can multiply our NCPs for pooled tests by $M/500$ and look up the corresponding tail probabilities for a noncentral chi-squared distribution.

The ratio of NCPs of two methods gives their relative efficiency. Conceptually, if one approach has a relative efficiency of 2 compared to another, the former approach requires only half as many triads as the latter to achieve the same power. The individual-LRT approach, however, requires h times the number of genotypes to achieve the same power if the relative efficiency is “1”.

We compared NCPs of three approaches: 1) triad members are genotyped individually and analyzed with a log-additive model (individual-LRT); 2) triad members are genotyped in pools of size 2 and analyzed with a log-additive model via EM (pool-LRT); 3) triad members are genotyped in pools of size 2 and analyzed with the difference-based Z^2 statistic (pool- Z^2).

Results

For all of the tests described below, NCPs were zero under the relevant null hypothesis, indicating consistency with any nominal Type I error rate. A simulation study testing only offspring genetic effects at the 0.05 level also produced nominal Type I error rates (Supplement S4). Under scenarios with population stratification, our proposed method maintained the correct Type I error rates as evidenced both by NCP calculations and by simulations (Supplement S5). We showed through simulation studies for both homogeneous and stratified scenarios (Supplements S4 and S5) that the power obtained based on NCP calculations agrees well with power based on simulations and that the relative risk estimates were unbiased when the risk model was correctly specified.

Test of effects of gene variants inherited by the offspring

In the absence of genotyping errors, the calculated relative efficiency of pool-LRT to individual-LRT was virtually 1 across different allele frequencies and that of pool- Z^2 to individual-LRT was imperceptibly lower (Figure 2A). Thus genotyping pools of size 2 achieved the virtually same power as genotyping individuals, with only half as many assays. When the error rate was moderate ($e_1=0.05$ and $e_2=0.048$) (Figure 2B) and high ($e_1=0.1$ and $e_2=0.091$) (Figure 2C), the relative efficiencies for both pooled methods dropped compared to individual genotyping. The geometric mean relative efficiency for pool-LRT versus individual-LRT across the 9 allele frequencies considered was 0.74 and 0.58 for moderate

and high error rates, respectively). The corresponding values for pool- Z^2 versus individual-LRT were 0.73 and 0.58. The power of pool-LRT was similar to or slightly better than that of pool- Z^2 across allele frequencies and error rates.

If only the cost of assays is of concern, one can fix the number of genotypes and compare the NCPs of pool-LRT to adjusted individual-LRT NCPs that correspond to using the same number of genotyping assays as the pool-LRT. This adjustment is calculated as $1/h$ times the NCPs of individual-LRT (shown in gray curves in Figures 2–5). Regardless of the error rate, pool-LRT outperformed individual-LRT in this sense in all scenarios considered except when the error rate was high and the allele frequency was at an extreme (0.1 or 0.9).

Test of maternal genetic effects on the offspring

The NCPs of maternal-effect tests showed similar patterns to those of offspring-effect tests (Figure 3). Pool-LRT performed similarly to pool- Z^2 test. The relative efficiency of pool-LRT to individual-LRT for maternal-effect tests was in general higher than that for offspring-effect tests. The relative increase was even more pronounced in the presence of genotyping errors. The geometric mean relative efficiencies across the 9 allele frequencies were 0.89 and 0.80 for the moderate and high error rate scenarios, respectively, under log-additive scenarios (Figure 5). For equal numbers of genotyping assays, pool-LRT was more efficient than individual-LRT across the allele frequencies and genotyping error rates considered.

Missing parent scenarios

As expected, power for both offspring and maternal effect tests using all three methods declined as the number of missing parents increased (Figures 4–5) but the relationship among the tests mirrored that seen without missing parents. For equal numbers of subjects, individual-LRT was generally more powerful than pool-LRT or pool- Z^2 . Pool-LRT was more powerful than pool- Z^2 . The geometric mean relative efficiencies compared to pool- Z^2 across the 9 allele frequencies were 1.08, 1.09 and 1.1 for no, moderate and high error rate scenarios in testing maternal effects, respectively. In addition, with some parents missing, the validity of pool- Z^2 rests on strong assumptions (e.g., when testing for offspring effects, one must assume no maternal effects exist) that are not required by pool-LRT. Violations of such assumptions inflated Type I error rates for pool- Z^2 but not for pool-LRT (data not shown). For equal numbers of genotyping assays, the relative efficiencies of pool-LRT to individual-LRT were similar to those under complete triad scenarios.

Results based on the alternative parameterization of errors (e_u , e_d) were similar (Supplement S3). We also evaluated the impact of an incorrectly specified error model via simulations where the error structure generating the data was more complex than our fitted error model (Supplement S7). Despite the misspecification, pool-LRT performed reasonably well: Type I error rates were slightly inflated; power was higher than that of pool- Z^2 though lower than individual-LRT; relative risk estimates showed some bias; however, error rates were badly over-estimated (Tables S7.3 and S7.4). We also calculated NCPs using a pool size of 3 for all the scenarios that we studied for a pool size of 2, and found similar relationships (online Supplement S3). We also evaluated the impact of relative risk on power and found that, as

expected, power improves with increasing relative risk for all methods considered. The relative efficiency between different methods was reasonably constant across different effect sizes (Supplement S6)

Discussion

Reducing the cost of genotyping for association studies, particularly for genome-wide studies, has prompted consideration of pooled DNA specimens for SNP genotyping. Several investigators have compared allele frequencies in pools of cases versus pools of controls, where “controls” might be unrelated individuals or, in family-based studies, parents or unaffected siblings (reviewed by Wang *et al.*, 2007). These proposals rely on measured allele frequencies from each pool and generally involve pooling large numbers (tens to hundreds) of individuals. Accurate DNA quantification and careful pool construction are essential to reduce measurement variance and to ensure equimolar amounts of DNA in a pool (Sham *et al.*, 2002). Estimated allele frequencies from pooled DNA have shown high reproducibility (Kirov *et al.*, 2006, Sham *et al.*, 2002, Meaburn *et al.*, 2006, Ozerov *et al.*, 2013, Uemoto *et al.*, 2012) and approaches that estimate allele frequencies in pools have been successfully used in many genome-wide association studies (Lu *et al.*, 2012 among others, Liu *et al.*, 2011, Janicki *et al.*, 2011, Krumbiegel *et al.*, 2011, Diergaarde *et al.*, 2010, Huang *et al.*, 2010, Tournas *et al.*, 2010, Craig *et al.*, 2009, Kirov *et al.*, 2009, Kawase *et al.*, 2008, Zaharieva *et al.*, 2008, Abraham *et al.*, 2008, Melquist *et al.*, 2007). Especially when working within tight budget constraints, pooling strategies offer effective ways to reduce genotyping costs while conserving limited specimens.

Lee (2005) proposed a useful approach to DNA pooling for case-parent studies. The idea is to recapitulate the essential triad data structure by forming pooling sets of h triads each and creating three DNA pools from each pooling set, a maternal pool, a paternal pool and an offspring pool. Compared to genotyping every individual, this strategy reduces the genotyping effort by a factor of $1/h$. In addition, while not considered by Lee, retaining the triad structure among the DNA pools allows exploration of maternal genetic effects or parent-of-origin effects.

Our proposed approach for the analysis of such pooled triad data, pool-LRT, maximizes the observed data likelihood by probabilistically disentangling the individual genotypes within the pools via the EM algorithm. Although pool-LRT employs a log-linear model and can thereby target a wide range of mechanisms of effect, we employed a log-additive risk model so we could compare our approach to pool- Z^2 , a statistic closely related to one proposed by Lee (2005) based on pooling-set-specific differences. Our pool- Z^2 differed from the related statistic proposed by Lee in the choice of the estimate of the variance in the denominator: we used the usual sample variance whereas Lee used an estimate that is unbiased under the null hypothesis that the expected value of the differences is zero. The variance that we used can never exceed that used by Lee so that, under alternatives, our pool- Z^2 has a larger NCP, hence power, than does Lee's test.

The statistical performance of these pooling approaches compared to analyses based on individual genotyping can be evaluated in two ways. First, one can consider holding the

number of triads constant for all analyses so the pooling approaches involve fewer genotyping assays; alternately, one can consider holding the number of genotyping assays constant so that the pooling approaches include more triads. In the former situation, when genotyping error is absent and pool size is two or three, our calculations showed that both these pooling approaches have about the same statistical efficiency for testing offspring genetic effects as the individual-LRT, a likelihood method based on individual genotyping, while substantially reducing the genotyping effort. Thus, for the pool sizes that we investigated, pooling *per se* imposes little or no loss of power. When genotyping errors are present, however, both pooling approaches suffer substantial loss of statistical efficiency compared to the individual-LRT, a phenomenon reported earlier for different tests and a different pooling strategy (Zou & Zhao, 2005). In the situation where the number of genotyping assays is held constant, with pool sizes of two or three, both pooling approaches outperform individual genotyping except at extreme allele frequencies. These general relationships of statistical efficiency for offspring effects among the pool-LRT, pool- Z^2 , and individual-LRT procedures were also seen for maternal effects. In addition, the relative efficiency for the pool-LRT and pool- Z^2 procedures compared with individual-LRT, in general, was higher for tests of maternal effects than for tests of offspring effects.

The similarity in performance for testing offspring effects among pool- Z^2 and the log-additive model versions of pool-LRT and individual-LRT may seem at first surprising but is understandable because of a close relationship among the difference scores D_k and the additive coding of offspring genotype in a log-linear model that includes parental mating types. Implicit in fitting the log-linear model, the additive coding is made orthogonal to the mating-type indicator variables. The orthogonalized additive coding turns out to be $D_k/2$. Consequently, the fitted log-linear model, and any resulting tests, are exactly the same whether the additive coding or $D_k/2$ is used as the offspring genotype variable. Using D_k instead of $D_k/2$ re-scales estimates of offspring effects but leaves tests unchanged.

Given the potential advantages of the pooling approaches, what characteristics distinguish the pool-LRT approach from its pool- Z^2 counterpart? In terms of statistical efficiency, neither has a clear advantage in the absence of genotyping error; but pool-LRT offers better power in the presence of genotyping errors, at least for the error models that we considered, and when some parental genotypes are missing. Important trade-offs between these approaches also lie in other areas. A principal advantage of the pool- Z^2 approach is the simplicity of the calculations involved compared to the more computationally intensive EM algorithm of the pool-LRT approach. Also, pool- Z^2 is applicable to pools of any size. Practicalities of the computationally intensive EM algorithm limit the pool-LRT approach to maximum pool sizes of two or three. A limitation of the pool- Z^2 approach is its reliance on additional and somewhat restrictive assumptions when parents are missing. A principal advantage of the pool-LRT approach is its ability to estimate genotype relative risk parameters and to test any hypothesis that is amenable to testing with case-parents data without those restrictive assumptions. For example, the pool-LRT approach can be used to study parent-of-origin effects. Although statistical power suffers more for parent-of-origin tests than for maternal or inherited effects, pooling may nevertheless be advantageous for studying parent of origin in a setting where genotyping is close to error-free (data not

shown). As technologies for genotyping continue to improve, this advantage will probably be realized. The pool-LRT could also be adapted to study gene-environment interactions using the ideas in Umbach and Weinberg (Umbach & Weinberg, 2000) and Weinberg and Umbach (Weinberg & Umbach, 2000) for an individually measured exposure or maternal-fetal interactions using ideas in Sinsheimer et al. (2003), though we have not yet investigated these possibilities. Further extensions may allow for pooled measurements of both genotype and a biomarker.

If one is concerned that a categorical exposure may be an effect modifier for a genetic effect, *e.g.*, sex of the offspring or maternal smoking status for oral clefting, then pooling sets should be formed in a way that matches on that factor. Creating pooling sets that are homogeneous as to subphenotype, *e.g.*, cleft palate only versus cleft lip with or without cleft palate, may also be useful by enabling analyses that assess etiologic heterogeneity in the effect of a variant allele.

A limitation that applies to DNA pooling strategies broadly is that statistical power suffers in the face of genotyping error, especially since genotyping errors are considered more likely when genotyping pools as opposed to individuals because of possible errors introduced in construction of pools (Earp *et al.*, 2011) and the inherent complexity of distinguishing among $2h+1$ possible genotypes. In addition, the impact of genotyping error is greater on pooled than individual genotyping (Zou & Zhao, 2004). Our pool-LRT analysis, which assumed that the assay would report the count of variant alleles for a pooled specimen, could in principle accommodate an arbitrary genotype misclassification matrix. We were unable to find information in the literature to guide the formulation of a realistic error model. Considerations of computation feasibility limited us to fitting a relatively simple two-parameter error models. When the fitted error model matched the error model that generated the data pool-LRT performed well. When the fitted error model misspecified the data-generating error model, we expected the performance of pool-LRT to degrade. Surprisingly, despite poor estimation of the error parameters under misspecification, estimation and testing of risk parameters were only moderately impacted. Nevertheless, having a realistic error model is important. A validation sub-sample where some pooling sets are measured both as pooled specimens and as individuals could help the analyst construct a realistic error model.

In the context of a GWAS using case-parent triads, one could carry out a pool-based screen in stage 1 and then follow up the top hits with models based on individually genotyped data. Stratification of stage 1 according to a measured exposure known to be related to risk, as discussed above, could also serve to provide additional statistical power for finding genetic effects and insight into gene-by-environment interaction (Kraft *et al.*, 2007).

Using a pooling strategy specifically designed for case-parents triads, we found that, when the number of genotyping assays is held constant, statistical procedures based on pooled genotypes (that consequently use more triads) generally outperform procedures based on individual genotypes. On the other hand, when the number of triads is held constant, procedures based on individual genotyping outperform those based on pooled genotyping (that consequently involve fewer assays), especially when genotyping errors are present.

Absent such errors, genotyping small pools imposes remarkably little loss of power compared to genotyping individuals.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences, under project numbers Z01 ES040007 and Z01 ES045002. We thank Drs Chia-Ling Kuo and Richard Morris for their careful review and valuable comments, the Computational Biology Core at NIEHS for facilitating our computationally intensive study.

References

- Abraham R, Moskvina V, Sims R, Hollingworth P, Morgan A, Georgieva L, Dowzell K, Cichon S, Hillmer AM, O'donovan MC, Williams J, Owen MJ, Kirov G. A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. *BMC medical genomics*. 2008; 1:44. [PubMed: 18823527]
- Agresti, A. *Categorical Data Analysis*. New York: John Wiley & Sons; 1990.
- Bader JS, Sham P. Family-based association tests for quantitative traits using pooled DNA. *Eur J Hum Genet*. 2002; 10:870–878. [PubMed: 12461696]
- Beckman KB, Abel KJ, Braun A, Halperin E. Using DNA pools for genotyping trios. *Nucleic Acids Res*. 2006; 34:e129. [PubMed: 17020923]
- Craig DW, Millis MP, Distefano JK. Genome-wide SNP genotyping study using pooled DNA to identify candidate markers mediating susceptibility to end-stage renal disease attributed to Type 1 diabetes. *Diabetic Med*. 2009; 26:1090–1098. [PubMed: 19929986]
- Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the *EM* Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1977; 39:1–38.
- Diergaarde B, Brand R, Lamb J, Cheong SY, Stello K, Barmada MM, Feingold E, Whitcomb DC. Pooling-based genome-wide association study implicates gamma-glutamyltransferase 1 (GGT1) gene in pancreatic carcinogenesis. *Pancreatology*. 2010; 10:194–200. [PubMed: 20484958]
- Earp MA, Rahmani M, Chew K, Brooks-Wilson A. Estimates of array and pool-construction variance for planning efficient DNA-pooling genome wide association studies. *BMC medical genomics*. 2011; 4:81. [PubMed: 22122996]
- Huang W, Kirkpatrick BW, Rosa GJ, Khatib H. A genome-wide association study using selective DNA pooling identifies candidate markers for fertility in Holstein cattle. *Animal genetics*. 2010; 41:570–578. [PubMed: 20394602]
- Janicki PK, Vealey R, Liu J, Escajeda J, Postula M, Welker K. Genomewide Association study using pooled DNA to identify candidate markers mediating susceptibility to postoperative nausea and vomiting. *Anesthesiology*. 2011; 115:54–64. [PubMed: 21694509]
- Kawase T, Nannya Y, Torikai H, Yamamoto G, Onizuka M, Morishima S, Tsujimura K, Miyamura K, Kodera Y, Morishima Y, Takahashi T, Kuzushima K, Ogawa S, Akatsuka Y. Identification of human minor histocompatibility antigens based on genetic association with highly parallel genotyping of pooled DNA. *Blood*. 2008; 111:3286–3294. [PubMed: 18178869]
- 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
- Kirov G, Zaharieva I, Georgieva L, Moskvina V, Nikolov I, Cichon S, Hillmer A, Toncheva D, Owen MJ, O'donovan MC. A genome-wide association study in 574 schizophrenia trios using DNA pooling. *Molecular psychiatry*. 2009; 14:796–803. [PubMed: 18332876]
- Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered*. 2007; 63:111–119. [PubMed: 17283440]

- Krumbiegel M, Pasutto F, Schlotzer-Schrehardt U, Uebe S, Zenkel M, Mardin CY, Weisschuh N, Paoli D, Gramer E, Becker C, Ekici AB, Weber BH, Nurnberg P, Kruse FE, Reis A. Genome-wide association study with DNA pooling identifies variants at CNTNAP2 associated with pseudoexfoliation syndrome. *Eur J Hum Genet.* 2011; 19:186–193. [PubMed: 20808326]
- Lee WC. A DNA pooling strategy for family-based association studies. *Cancer Epidemiol Biomarkers Prev.* 2005; 14:958–962. [PubMed: 15824170]
- Liu L, Li J, Yao J, Yu J, Zhang J, Ning Q, Wen Z, Yang D, He Y, Kong X, Song Q, Chen M, Yang H, Liu Q, Li S, Lin J. A genome-wide association study with DNA pooling identifies the variant rs11866328 in the GRIN2A gene that affects disease progression of chronic HBV infection. *Viral immunology.* 2011; 24:397–402. [PubMed: 22004137]
- Lu Y, Chen X, Beesley J, Johnatty SE, Defazio A, Lambrechts S, Lambrechts D, Despierre E, Vergotes I, Chang-Claude J, Hein R, Nickels S, Wang-Gohrke S, Dork T, Durst M, Anttonenkov N, Bogdanova N, Goodman MT, Lurie G, Wilkens LR, Carney ME, Butzow R, Nevanlinna H, Heikkinen T, Leminen A, Kiemeny LA, Massuger LF, Van Altena AM, Aben KK, Kjaer SK, Hogdall E, Jensen A, Brooks-Wilson A, Le N, Cook L, Earp M, Kelemen L, Easton D, Pharoah P, Song H, Tyrer J, Ramus S, Menon U, Gentry-Maharaj A, Gayther SA, Bandera EV, Olson SH, Orlow I, Rodriguez-Rodriguez L, Macgregor S, Chenevix-Trench G. Genome-wide association study for ovarian cancer susceptibility using pooled DNA. *Twin research and human genetics : the official journal of the International Society for Twin Studies.* 2012; 15:615–623. [PubMed: 22794196]
- Meaburn E, Butcher LM, Schalkwyk LC, Plomin R. Genotyping pooled DNA using 100K SNP microarrays: a step towards genomewide association scans. *Nucleic Acids Res.* 2006; 34
- Melquist S, Craig DW, Huentelman MJ, Crook R, Pearson JV, Baker M, Zismann VL, Gass J, Adamson J, Szelinger S, Corneveaux J, Cannon A, Coon KD, Lincoln S, Adler C, Tuite P, Calne DB, Bigio EH, Uitti RJ, Wszolek ZK, Golbe LI, Caselli RJ, Graff-Radford N, Litvan I, Farrer MJ, Dickson DW, Hutton M, Stephan DA. Identification of a novel risk locus for progressive supranuclear palsy by a pooled genomewide scan of 500,288 single-nucleotide polymorphisms. *Am J Hum Genet.* 2007; 80:769–778. [PubMed: 17357082]
- Ozerov M, Vasemagi A, Wennevik V, Niemela E, Prusov S, Kent M, Vaha JP. Cost-effective genome-wide estimation of allele frequencies from pooled DNA in Atlantic salmon (*Salmo salar* L.). *Bmc Genomics.* 2013; 14:12. [PubMed: 23324082]
- Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res.* 1998; 8:1273–1288. [PubMed: 9872982]
- Sham P, Bader JS, Craig I, O'donovan M, Owen M. DNA Pooling: a tool for large-scale association studies. *Nat Rev Genet.* 2002; 3:862–871. [PubMed: 12415316]
- Sinsheimer JS, Palmer CG, Woodward JA. Detecting genotype combinations that increase risk for disease: maternal-fetal genotype incompatibility test. *Genet Epidemiol.* 2003; 24:1–13. [PubMed: 12508251]
- Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet.* 1993; 52:506–516. [PubMed: 8447318]
- Tournas A, Mfuna L, Bosse Y, Filali-Mouhim A, Grenier JP, Desrosiers M. A pooling-based genome-wide association study implicates the p73 gene in chronic rhinosinusitis. *Journal of otolaryngology - head & neck surgery = Le Journal d'oto-rhino-laryngologie et de chirurgie cervico-faciale.* 2010; 39:188–195.
- Uemoto Y, Sasago N, Abe T, Okada H, Maruoka H, Nakajima H, Shoji N, Maruyama S, Kobayashi N, Mannen H, Kobayashi E. Practical capability of a DNA pool-based genome-wide association study using BovineSNP50 array in a cattle population. *Animal science journal = Nihon chikusan Gakkaiho.* 2012; 83:719–726. [PubMed: 23126324]
- Umbach DM, Weinberg CR. The use of case-parent triads to study joint effects of genotype and exposure. *Am J Hum Genet.* 2000; 66:251–261. [PubMed: 10631155]
- Wang, J.; Zou, G.; Zhao, H. DNA pooling: methods and applications in association studies. In: Deng, HW.; Shen, H.; Liu, Y., editors. *Current Topics in Human Genetics: Studies in Complex Diseases.*

- 1 ed. Singapore: World Scientific Publishing Company; 2007. Current Topics in Human Genetics: Studies in Complex Diseases
- Weinberg CR. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am J Hum Genet.* 1999; 65:229–235. [PubMed: 10364536]
- Weinberg CR, Umbach DM. Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *Am J Epidemiol.* 2000; 152:197–203. [PubMed: 10933265]
- Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet.* 1998; 62:969–978. [PubMed: 9529360]
- Wilcox AJ, Weinberg CR, Lie RT. Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads". *Am J Epidemiol.* 1998; 148:893–901. [PubMed: 9801020]
- Zaharieva I, Georgieva L, Nikolov I, Kirov G, Owen MJ, O'donovan MC, Toncheva D. Association study in the 5q31-32 linkage region for schizophrenia using pooled DNA genotyping. *Bmc Psychiatry.* 2008; 8
- Zou G, Zhao H. The impacts of errors in individual genotyping and DNA pooling on association studies. *Genet Epidemiol.* 2004; 26:1–10. [PubMed: 14691952]
- Zou G, Zhao H. Family-based association tests for different family structures using pooled DNA. *Ann Hum Genet.* 2005; 69:429–442. [PubMed: 15996171]

Observed pooling-set genotype (T_k)	Pooling-set genotypes consistent with the observed T_k ($T_k^E \in \mathbb{T}(E, T_k)$)	Probability of observed given $T_k^E = T_k^*$ ($\Pr(T_k \mid T_k^E = T_k^*)$)	Compatible triad configurations ($C \in \mathbb{C}(T_k^E),$)	Weight (w_s)
(4, 4, 4)	1. (4, 4, 4)	1. $(1 - e_1)^3$	1. ((2, 2, 2), (2, 2, 2))	w_1
	2. (4, 4, 3)	2. $(e_2)(1 - e_1)^2$	None	
	3. (4, 3, 4)	3. $(e_2)(1 - e_1)^2$	3.1 ((2, 1, 2), (2, 2, 2))	w_2
	4. (3, 4, 4)	4. $(e_2)(1 - e_1)^2$	3.2 ((2, 2, 2), (2, 1, 2))	w_3
	5. (4, 3, 3)	5. $(e_2)^2(1 - e_1)$	4.1 ((1, 2, 2), (2, 2, 2))	w_4
	6. (3, 4, 3)	6. $(e_2)^2(1 - e_1)$	4.2 ((2, 2, 2), (1, 2, 2))	w_5
	7. (3, 3, 4)	7. $(e_2)^2(1 - e_1)$	5.1 ((2, 2, 2), (2, 1, 1))	w_6
	8. (3, 3, 3)	8. $(e_2)^3$	5.2 ((2, 1, 1), (2, 2, 2))	w_7
			6.1 ((1, 2, 1), (2, 2, 2))	w_8
			6.2 ((2, 2, 2), (1, 2, 1))	w_9
			7.1 ((1, 1, 2), (2, 2, 2))	w_{10}
			7.1 ((2, 2, 2), (1, 1, 2))	w_{11}
			7.3 ((1, 2, 2), (2, 1, 2))	w_{12}
			7.4 ((2, 1, 2), (1, 2, 2))	w_{13}
			8.1 ((1, 1, 1), (2, 2, 2))	w_{14}
			8.2 ((2, 2, 2), (1, 1, 1))	w_{15}
			8.3 ((1, 2, 1), (2, 1, 2))	w_{16}
			8.4 ((2, 1, 2), (1, 2, 1))	w_{17}
			8.5 ((1, 2, 2), (2, 1, 1))	w_{18}
			8.6 ((2, 1, 1), (1, 2, 2))	w_{19}

Figure 1.

An example of compatible triad configurations $\mathbb{C}(T_k^E)$ under our error model for a pool size of 2 when $T_k = (4, 4, 4)$. Weights of all configurations compatible with T_k sum to 1.

Figure 2.

Noncentrality parameter and power for tests of offspring genetic effects. All designs used 1000 complete triads under the risk scenario: $R_1=1.2$, $R_2=1.44$, $S_1=1$ and $S_2=1$. Vertical axes: left, the chi-squared noncentrality parameter for a 1-df likelihood ratio test; right, power at $\alpha=0.05$ (horizontal lines mark selected power levels). Horizontal axis shows the allele frequency ranging from 0.1 to 0.9. Panels: (A) no genotyping errors, *i.e.*, $e_1 = 0$ and $e_2=0$; (B) moderate genotyping errors, *i.e.*, $e_1 = 0.05$ and $e_2 = 0.048$; (C) high genotyping errors, *i.e.*, $e_1 = 0.1$ and $e_2 = 0.091$. Curves: **solid**, individual-LRT (triads individually genotyped and analyzed using a log-additive model); **dash-dot**, pool-LRT (pooled samples with a pool size of 2 analyzed using a log-additive model via EM); **dashed**, pool- Z^2 (pooled samples with a pool size of 2 analyzed using the difference-based approach). The grey line shows the NCPs when $\frac{1}{2}$ of triads are individually genotyped (same genotyping effort as with pooled triads).

Figure 3.

Noncentrality parameter and power for tests of maternal genetic effects. All designs used 1000 complete triads under the risk scenario: $R_1 = 1$, $R_2 = 1$, $S_1 = 1.2$ and $S_2 = 1.44$. Vertical axes: left, the chi-squared noncentrality parameter for a 1-df likelihood ratio test; right, power at $\alpha=0.05$ (horizontal lines mark selected power levels). Horizontal axis shows the allele frequency ranging from 0.1 to 0.9. Panels: (A) no genotyping errors, *i.e.*, $e_1 = 0$ and $e_2 = 0$; (B) moderate genotyping errors, *i.e.*, $e_1 = 0.05$ and $e_2 = 0.048$; (C) high genotyping errors, *i.e.*, $e_1 = 0.1$ and $e_2 = 0.091$. Curves: **solid**, individual-LRT (triads individually genotyped and analyzed using a log-additive model); **dash-dot**, pool-LRT (pooled samples with a pool size of 2 analyzed using a log-additive model via EM); **dashed**, pool- Z^2 (pooled samples with a pool size of 2 analyzed using the difference-based approach). The grey line shows the NCPs when $\frac{1}{2}$ of triads are individually genotyped (same genotyping effort as with pooled triads).

Figure 4.

Noncentrality parameter and power for tests of offspring genetic effects when 50% families have father's genotype missing. All designs used 1000 complete triads under the risk scenario: $R_1=1.2$, $R_2=1.44$, $S_1=1$ and $S_2=1$. Vertical axes: left, the chi-squared noncentrality parameter for a 1-df likelihood ratio test; right, power at $\alpha=0.05$ (horizontal lines mark selected power levels). Horizontal axis shows the allele frequency ranging from 0.1 to 0.9. Panels: (A) no genotyping errors, *i.e.*, $e_1 = 0$ and $e_2 = 0$; (B) moderate genotyping errors, *i.e.*, $e_1 = 0.05$ and $e_2 = 0.048$; (C) high genotyping errors, *i.e.*, $e_1 = 0.1$ and $e_2 = 0.091$. Curves: **solid**, individual-LRT (triads individually genotyped and analyzed using a log-additive model); **dash-dot**, pool-LRT (pooled samples with a pool size of 2 analyzed using a log-additive model via EM); **dashed**, pool- Z^2 (pooled samples with a pool size of 2 analyzed using the difference-based approach). The grey line shows the NCPs when $\frac{1}{2}$ of triads are individually genotyped (same genotyping effort as with pooled triads).

Figure 5.

Noncentrality parameter and power for tests of maternal genetic effects when 50% families have father's genotype missing. All designs used 1000 complete triads under the risk scenario: $R_1=1$, $R_2=1$, $S_1=1.2$ and $S_2=1.44$. Vertical axes: left, the chi-squared noncentrality parameter for a 1-df likelihood ratio test; right, power at $\alpha=0.05$ (horizontal lines mark selected power levels). Horizontal axis shows the allele frequency ranging from 0.1 to 0.9. Panels: (A) no genotyping errors, *i.e.*, $e_1=0$ and $e_2=0$; (B) moderate genotyping errors, *i.e.*, $e_1=0.05$ and $e_2=0.048$; (C) high genotyping errors, *i.e.*, $e_1=0.1$ and $e_2=0.091$. Curves: **solid**, individual-LRT (triads individually genotyped and analyzed using a log-additive model); **dash-dot**, pool- LRT (pooled samples with a pool size of 2 analyzed using a log-additive model via EM); **dash**, pool- Z^2 (pooled samples with a pool size of 2 analyzed using the difference-based approach). The grey line shows the NCPs when $\frac{1}{2}$ of triads are individually genotyped (same genotyping effort as with pooled triads).