Statistical Applications in Genetics and Molecular Biology

Volume 8, *Issue* 1 2009 *Article* 24

Incorporating Duplicate Genotype Data into Linear Trend Tests of Genetic Association: Methods and Cost-Effectiveness

Bryce Borchers[*] Marshall Brown[†] Brian McLellan[‡]
Airat Bekmetjev[**] Nathan L. Tintle[††]

[*] Rose-Hulman Institute of Technology, borcheb1@rose-hulman.edu

[†] Seattle Pacific University, brownm2@spu.edu

[‡] Hope College, brian.mclellan@hope.edu

[**] Hope College, bekmetjev@hope.edu

[††] Hope College, tintle@hope.edu

Incorporating Duplicate Genotype Data into Linear Trend Tests of Genetic Association: Methods and Cost-Effectiveness*

Bryce Borchers, Marshall Brown, Brian McLellan, Airat Bekmetjev, and Nathan L. Tintle

Abstract

The genome-wide association (GWA) study is an increasingly popular way to attempt to identify the causal variants in human disease. Duplicate genotyping (or re-genotyping) a portion of the samples in a GWA study is common, though it is typical for these data to be ignored in subsequent tests of genetic association. We demonstrate a method for including duplicate genotype data in linear trend tests of genetic association which yields increased power. We also consider the cost-effectiveness of collecting duplicate genotype data and find that when the relative cost of genotyping to phenotyping and sample acquisition costs is less than or equal to the genotyping error rate it is more powerful to duplicate genotype the entire sample instead of spending the same money to increase the sample size. Duplicate genotyping is particularly cost-effective when SNP minor allele frequencies are low. Practical advice for the implementation of duplicate genotyping is provided. Free software is provided to compute asymptotic and permutation based tests of association using duplicate genotype data as well as to aid in the duplicate genotyping design decision.

KEYWORDS: re-genotype, genome-wide association, genotyping error, genotype error, SNP

*Bryce Borchers, Marshall Brown and Brian McLellan contributed equally to this work. This project was funded in part by a grant from the National Institutes of Health, R15-HG004543. The content is solely the responsibility of the authors and does not necessarily represent the official view of the National Human Genome Research Institute or the National Institutes of Health. Additionally, this project received support from the Tanis Fund for Statistics Research.

Introduction

Genetic tests of association often utilize case-control study designs in order to identify possible genetic factors contributing to the etiology of a complex disease (Amos 2007, Sasieni 1997). Examining the whole genome simultaneously through genome-wide association (GWA) studies has become an increasingly popular and effective method of determining genetic association. While high costs of GWA studies are still a limiting factor, they continue to become more economically plausible with advances in technology that identify single nucleotide polymorphism (SNP) genotypes at decreasing costs (Amos 2007).

Despite these technological advances, the misclassification of genotypes by SNP technology (genotyping errors) remains a persistent issue. Genotyping error rates are low in many instances (~0.1-0.2% or lower; Saunders et al. 2007, Tintle et al. 2005, Fridley et al. 2008, Heid et al. 2008, Pompanon et al. 2005). However, these error rates are not uniform across all SNPs and some SNPs have measurably larger genotyping error rates (Pompanon et al. 2005). The impact of genotyping errors on case-control tests of genotype-phenotype association is well known. Specifically, non-differential errors (genotyping error rates are the same regardless of phenotype) have no effect on type I error, but do cause inflated type II error (i.e. reduce power) (Gordon & Ott 2001, Gordon et al. 2002, Ahn et al. 2007). Genotyping errors are particularly detrimental to power when the minor SNP allele frequency is low (Gordon et al. 2002, Ahn et al. 2007, Kang et al. 2004, Gordon & Finch 2005).

In addition to laboratory and technology-based approaches to reducing genotyping errors, which seek to address errors at their source, some have proposed the consideration of genotyping errors when designing the study. For example, double sampling (Gordon et al. 2004, Gordon et al. 2007) uses a perfect genotype mechanism (like gene sequencing) on a subset of the sample. Another recent paper discusses how to incorporate genotyping errors when optimizing a two-stage design (Zuo et al. 2008). A third approach involves replicate genotyping (Fridley et al. 2008, Rice & Holmans 2003, Tintle et al. 2007, Lai et al. 2007, Bonin et al. 2004), which means genotyping a random subset of individuals in the sample two or more times, instead of only once.

Duplicate genotyping has been proposed by many for quality control reasons (e.g. Rice & Holmans 2003, Bonin et al. 2004) and it is now a fairly common practice (Tintle et al. 2005, Fridley et al. 2008). Traditionally, duplicate genotype data were ignored in the subsequent statistical analyses. The data were simply used as an initial assessment of data quality. Recently, however, a method was proposed to incorporate duplicate genotype data in standard χ_2^2 tests of genotype-phenotype association on 2x3 tables (Tintle et al. 2007).

Subsequently, Tintle et al. (to appear) demonstrated the cost-effectiveness of duplicate genotyping (i.e. more power) for use in χ_2^2 tests when genotyping costs are low relative to phenotyping and sample acquisition costs. It was found that, as a general rule, duplicate genotyping the entire sample increases power when relative genotype to phenotype/sample acquisition costs don't exceed the genotyping error rate. Additionally, when the minor SNP allele frequency is low, duplicate genotyping the entire sample can be cost-effective even when relative costs are greater than the genotyping error rate.

The linear trend test of association (LTT), first proposed by Cochran (1954) and Armitage (1955), has been suggested by many (Sasieni 1997, Slager & Schaid 2001, Freidlin et al. 2002, Zheng et al. 2003, Zheng & Gastwirth 2006) as a method for analyzing SNP genotype data since it can incorporate information about the disease mode of inheritance, and thus increase statistical power by narrowing the focus of the alternative hypothesis. Recently, Ahn et al. (2007) demonstrated the impact of genotyping errors on the LTT. Also, Gordon et al. (2007) demonstrated how to use the LTT when double sample data are collected. In this paper, we demonstrate how to include duplicate genotype data in a LTT. We also explore the utility of including duplicate genotype data in subsequent tests of association if they have been collected for quality control reasons. Lastly, we evaluate the cost-effectiveness of designing a study to collect duplicate genotype data for analysis with the LTT.

Methods

Sampling Strategy

We consider a sampling strategy where a fraction of the entire sample, r ($r \in [0,1]$), is randomly selected to be genotyped exactly twice, while the remaining fraction of the sample, *(1-r),* is genotyped exactly once. We assume that all samples have been phenotyped as either a "case" or a "control."

Genotyping Error Assumptions

1. Let $\varepsilon_{i,j}$ be the probability of an individual of genotype i being classified as genotype j. Following the error model of Douglas et al. (2002), we assume that $\varepsilon_{1,2} = \varepsilon_{2,1} = \varepsilon_{2,3} = \varepsilon_{3,2}$ and $\varepsilon_{1,3} = \varepsilon_{3,1} = 0$.

2. We assume non-differential genotyping errors, meaning that the probability of genotyping errors is the same for each individual in the sample, regardless of case or control status.

3. We assume that genotyping error probabilities are independent and remain constant from the first to second genotyping. Specifically, we mean that the probability of a genotyping error does not change for an individual's second genotyping, and is not dependent upon whether they were incorrectly genotyped the first time.

Notation

$\delta_m =$ the frequency of allele *m* at the SNP marker. In this paper we assume the SNP is bi-allelic, and, thus, *m=U,V*. We also assume that the SNP marker allele associated with the disease is allele *2*.

$\zeta_n =$ the frequency of risk allele *n* at the disease locus. In this paper we assume the disease locus is bi-allelic and we denote the risk allele as *B* and the non-risk allele as *A*. Thus, *n=A,B*.

$h_{mn} =$ the frequency of the *mn* haplotype; that is, the frequency of having both the *m* allele at the SNP marker and risk allele *n* at the disease locus. Thus, $\sum_{m=U,V} \sum_{n=A,B} h_{mn} = 1$.

$D =$ the unstandardized measure of linkage disequilibrium between the SNP marker, *V*, and the disease risk allele, *B*. Thus,
$D = h_{VB} - \delta_V \zeta_B$.

$r^2 =$ the measure of the correlation between the SNP marker and the disease risk allele $= D^2 / (\delta_U \delta_V \zeta_A \zeta_B)$.

$\rho = r^2 / \max(r^2) =$ a measure of the correlation of SNP allele *V* and disease risk allele *B* as a fraction of their maximum possible correlation. As pointed out by Amos (2007), max(r^2)<1 unless $\delta_V = \zeta_B$. We also note that for any values of δ_V and ζ_B, max(r^2) is attained when $D' = 1$, where $D' = D / \min(\delta_U \zeta_B, \delta_V \zeta_A)$.

$\phi =$ the disease prevalence in the population.

$f_{j_1 j_2}$ = the penetrance of the disease given genotype j_1j_2 at the disease locus. Thus, f_{BB} is the probability someone who is BB at the disease locus (homozygote for the risk allele) has the disease, f_{AB} is the probability someone who is AB at the disease locus (heterozygote for the risk allele) has the disease, and f_{AA} is the probability someone who is AA at the disease locus (homozygote for the non-risk allele) has the disease.

$\gamma =$ a general relative risk of disease parameter which is used to compute genotype specific relative risks (γ_{BB} and γ_{AB}) in ways that are dependent upon the mode of inheritance of the disease (dominant, additive, recessive).

p_i = the probability of genotype *i* in the cases, *i*=1,2,3

q_i = the probability of genotype *i* in the controls, *i*=1,2,3

p_i^* = the probability of observing genotype *i* in the cases assuming there are genotyping errors, and the sample is genotyped exactly once *i*=1,2,3

q_i^* = the probability of observing genotype *i* in the controls assuming there are genotyping errors, and the sample is genotyped exactly once *i*=1,2,3

p_{ij}^* = the probability of observing genotype *i* once and genotype *j* once in the cases assuming there are genotyping errors, and the sample is genotyped exactly twice *i*=1,2,3, *j*=1,2,3 and $i \le j$.

q_{ij}^* = the probability of observing genotype *i* once and genotype *j* once in the controls assuming there are genotyping errors, and the sample is genotyped exactly twice *i*=1,2, 3, *j*=1,2,3 and $i \le j$.

T = the total number of cases

S = the total number of controls

$N = T + S$ = the total sample size

$k = S/T$ = ratio of controls to cases

c = the relative cost of genotyping to phenotyping/sample acquisition

Contingency Tables for a Study with Duplicate Genotype Data

When a fraction *(r)* of the sample has been duplicate genotyped, and the SNP marker under consideration has three possible genotypes (1=*UU*, 2=*UV* and 3=*VV*), data can be summarized into two tables, as shown in *Tables 1a* and *1b*. We assume that an equal fraction of both cases and controls has been duplicate genotyped.

Table 1a. Single genotyped data

Genotype	1	2	3	Total
Cases	t_1	t_2	t_3	$T(1-r)$
Controls	s_1	s_2	s_3	$S(1-r)$
Total	n_1	n_2	n_3	$N(1-r)$

Table 1b. Duplicate genotyped data

Genotype	(11)	(12)	(13)	(22)	(23)	(33)	Total
Cases	t_{11}	t_{12}	t_{13}	t_{22}	t_{23}	t_{33}	Tr
Controls	s_{11}	s_{12}	s_{13}	s_{22}	s_{23}	s_{33}	Sr
Total	n_{11}	n_{12}	n_{13}	n_{22}	n_{23}	n_{33}	Nr

Using a weighting strategy for duplicate genotype data presented by Tintle et al. (2007), *Tables 1a* and *1b* can be combined into a single table (*Table 1c*) as follows:

$$t'_i = t_i + t_{ii} + 0.5(t_{ij}) + 0.5(t_{ik}) \tag{1}$$

where $i \neq j$ and $i \neq k$, with a similar equation for the controls. As in shown in Tintle et al. (2007), using equal weights (0.5) for the inconsistently identified individuals is optimal.

Table 1c. Combined data contingency table

Genotype	1	2	3	Total
Cases	t'_1	t'_2	t'_3	T
Controls	s'_1	s'_2	s'_3	S
Total	n'_1	n'_2	n'_3	N

Disease Modes of Inheritance

We consider three disease modes of inheritance (MOI): Dominant $(\gamma_{BB} = \gamma_{AB} = \gamma)$, Additive$(\gamma_{AB} = \gamma,\ \gamma_{BB} = 2\gamma - 1)$, and Recessive ($\gamma_{BB} = \gamma$, $\gamma_{AB} = 1$), where γ_{BB} = the relative risk of disease for a participant with two copies of the risk allele (f_{BB}/f_{AA}) and γ_{AB} = the relative risk of disease for a participant with one copy of the risk allele (f_{AB}/f_{AA}). Note that when $\gamma_{BB} = \gamma_{AB} = 1$ then the null hypothesis of no association between genotype and disease is true.

Linear Trend Test

As noted earlier, the linear trend test (LTT) is a powerful choice for the analysis of case-control studies of genetic association because of the ability to include information about the disease mode of inheritance (Sasieni 1997, Slager & Schaid 2001, Zheng & Gastwirth 2006). The traditional LTT statistic is U/σ_U where U is a statistic based on the disease mode of inheritance and the observed cell counts in the 2x3 contingency table (e.g. *Table 1a*), and σ_U is estimated based on the observed cell counts in the same table. In this paper, we extend the traditional version of the test to be able to include duplicate genotype data, proposing the *LTT_d (see Results:Finding the LTT statistic*). In short, the *LTT_d* uses the strategy proposed by Tintle et al. (2007) to place individuals who have been inconsistently duplicate genotyped to each of the two genotypes to which they have been genotyped (see Equation (*1*)). This strategy, however, means that the resulting contingency table of phenotype-genotype (*Table 1c*) no longer has a multinomial distribution due to increased covariance between cells, requiring the introduction of the *LTT_d.*

In developing the *LTT_d*, we also address the issue of bias in the σ_U estimate. Freidlin et al. (2002) demonstrated that the method of estimating σ_U as considered by Slager and Schaid (2001) was biased and thus provided invalid results. Zheng and Gastwirth (2006) consider two alternatives to the Slager and Schaid approach which they call “case-control” (cc) and “control” (c). The Slager and Schaid method estimates σ_U assuming that the null hypothesis of no genotype-phenotype association is true. The *cc* method estimates σ_U without the restriction of the null hypothesis being true, whereas the *c* method is similar but only uses the sample of controls. Zheng and Gastwirth find, and we confirmed in our own attempts to implement the method, that the *c* method increases the type I error in some cases (results not shown). Thus, we choose to base our results only on the *cc* method.

Simulation Study

To confirm that the empirical distribution of the LTT_d (Derived later, see *Results: Finding the* LTT_d *statistic*) follows the theoretical asymptotic distribution (χ_1^2) for practical sample sizes we conducted a simulation study (see *Table 2* for parameters and values used).

Table 2. Parameter values for the simulation study

Parameter	**Values**
δ_V	0.05, 0.20, 0.50
ζ_B	0.05, 0.20, 0.50
ρ	0.80, 1.0
ϕ	0.025, 0.10
ε	0.001, 0.01, 0.03
γ	1.00, 1.25, 2.00
r	0, 0.5, 1.0
N	1000, 5000
k	1.5, 2/3, 1.0
Disease MOI	Dominant, Additive, Recessive

We examined all possible combinations of parameter values and so a total of 17,496 settings were evaluated. The simulation study was conducted as follows:

Step 1. For given values of δ_V, ζ_B, ρ, ϕ, γ, and the disease MOI the true genotype probabilities (p_i and q_i, i=1,2,3) were computed. See Ahn et al. (2007) for details.

Step 2. The true genotype probabilities (p_i and q_i, i=1,2,3) were then adjusted to reflect the genotype error rate (ε), yielding p_{ij}^*, q_{ij}^*, p_i^* and q_i^*. See Tintle et al. (2007) for details.

Step 3. For given values of k, r, and n, and the observed single and duplicate genotyping probabilities ($p_{ij}^*, q_{ij}^*, p_i^*$ and q_i^*) found in step 2, entries into *Tables 1a* and *1b* were randomly simulated. For each combination of parameter values in *Table 2*, 2,000 random tables were simulated. In cases where γ=1 (null hypothesis is true), the type I error rate was analyzed by comparing the nominal significant level α (we examined 0.05, 0.005, and 0.0002) with the empirical α level. In cases where γ=1.25 or γ=2.00 (i.e. the alternative hypothesis is true), the empirical power was compared to the theoretical power (see equation *(A3)* in the *Appendix*).

Cost-effectiveness Computational Study

We completed a computational study comparing theoretical power values for different values of *r* (duplicate genotyping percentage), *c* (relative genotyping costs) and other parameters. *Table 3* shows the settings used for this study. We examined all 10,368 possible combinations of parameter values based on *Table 3*.

Table 3. Parameters and values for the computational study

Parameter	Values
δ_V	0.05, 0.20, 0.50
ζ_B	0.05, 0.20, 0.50
ρ	0.80, 1.0
ϕ	0.025, 0.10
ε	0.001, 0.01, 0.03
γ	1.25, 2.00
r	0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
c	0, 0.001,0.005,0.01,0.02,0.05,0.10,0.50
Power if there were no genotyping errors and no duplicates were collected	0.80, 0.95
k	1.5, 2/3, 1.0
Disease MOI	Dominant, Additive, Recessive

The computational study was carried out as follows:

Step 1. Assuming there are no genotyping errors, for given values of δ_V, ζ_B, ρ, ϕ, γ and the disease type, the genotype probabilities were computed as if no duplicates were obtained. These values were then used to find the sample size needed (N_0) to yield the specified power level (80% or 95%).

Step 2. Find the budget (*B*) needed to conduct the study if no duplicates as: $B=(1+c)N_0$, where *c* is the genotyping cost per person relative to phenotyping/acquisition cost.

Step 3. Assuming there is duplicate genotyping (*r*>0), the sample size that can be obtained for the same budget, *B*, is found as $N_r = \frac{B}{1+c(1+r)}$. N_r can then be used in the power computation formula *(A3)* in the *Appendix*, to find the power using duplicate genotyping for that sample size. Then we find the optimal value of *r* that yields the largest power of the test. All computations used α=0.0002.

Results

Finding the LTT_d Statistic

Zheng and Gastwirth (2006) present a test statistic for the LTT as $Z = \frac{U}{\sqrt{V}}$, where $U = \sum_i x_i \left(\frac{S}{N} t_i - \frac{T}{N} s_i \right)$ and *V* is an estimate of the variance of *U*. Tintle et al. (2007) showed that by using the allocation strategy *(Equation (1)) Tables 1a* and *1c* estimate the same quantities. Thus, the numerator of the Zheng and Gastwirth *Z* statistic becomes:

$$U_d = \sum_i x_i \left(\frac{S}{N} t'_i - \frac{T}{N} s'_i \right) \tag{2}$$

According to the Central Limit Theorem, *Table 1c* has an approximately multivariate normal distribution (see also Tintle et al. 2007). The expected value of U_d under the null hypothesis ($p^*_i = q^*_i$ for all *i*) is zero (see Equation *(A1)* in the *Appendix*). Thus,

$$LTT_d = \frac{U_d}{\sqrt{Var(U_d)}} \tag{3}$$

has a standard normal distribution and, therefore, $(LTT_d)^2$ has a χ^2_1 distribution. Following the results of Zheng and Gastwirth (2006) the expression for $Var(U_d)$ follows from *(A2; Appendix)*, using $p^*_i = \frac{r_i}{R_s}, q^*_i = \frac{s_i}{S_s}$ and $p^*_{ij} = \frac{r_{ij}}{R_d}, q^*_{ij} = \frac{s_{ij}}{S_d}$. Equation *(A2)* also accounts for additional covariance between cells in *Table 1c* from using the allocation strategy.

Simulation Results for LTT_d

As described earlier (*Methods: Simulation study*), a simulation study was conducted to ensure that nominal type I and type II error rates obtained using the asymptotic theory of the LTT_d were maintained in practice. First we consider the distribution of LTT_d if the null hypothesis is true ($\gamma = 1$) and then the distribution of LTT_d if the alternative hypothesis is true ($\gamma \neq 1$).

Simulation Results for LTT_d under the Null Hypothesis

For each combination of parameter values, a 99% confidence interval was found for the empirical α. For both the dominant and additive models, nominal type I error rates were maintained empirically regardless of sample size since an expected number of simulation settings had a 99% confidence interval on the empirical α that did not contain the nominal α (1.2% and 1.2% for dominant and additive, respectively, for α=0.05, 1.2% and 1.3% for the α=0.005 level and 1.1% and 0.7% for the α=0.0002 level). Nominal type I error rates were maintained empirically for the recessive model as long as the minimum cell count in *Table 1c* was at least 5 (detailed results not shown).

Simulation Results for LTT_d under the Alternative Hypothesis

The LTT_d statistic generally gives comparable theoretical and empirical power values across all simulation settings for the additive and dominant models as long as expected cell counts in *Table 1c* are at least 5. For each combination of parameter values, a 99% confidence interval was placed on the empirical power. For both the dominant and additive models when the minimum cell count in *Table 3* was at least 5, an expected number of simulation settings had a 99% confidence interval on the empirical power that did not contain the theoretical power (0.9% and 1.2% for dominant and additive, respectively, for α=0.05, 1.3% and 0.9% for α=0.005 level and 1.7% and 1.3% for the α=0.0002 level). When the minimum cell count was less than 5 in the dominant and additive models, the empirical power was often still very close to theoretical power (results not shown). The recessive model with small δ_V showed significant differences between theoretical and empirical power (detailed results not shown), though theoretical power and empirical power were similar for larger values of δ_V.

Recommendations for Use of a Permutation Test

Based on the simulation study, differences in theoretical and empirical type I and type II errors are possible when the recessive disease model is used, in cases where at least one cell count in the grouped table is less than 5, or in cases where the total sample is less than 1,000 individuals. In these cases we recommend estimating p-values for the LTT_d by permuting phenotype status instead of using the asymptotic theory provided above. A permutation based p-value is available in our software (see *Results: Software*).

Example

In Tintle et al. (2007), duplicate genotype data from a case-control study on bi-polar disorder was presented for a SNP with inconsistently genotyped individuals where all individuals were duplicate genotyped. We present this data here (*Table 4*) in the form of *Table 1b* using a linear trend test for analysis to demonstrate the utility of the methods just developed.

Table 4. Duplicate genotype data from a study of bi-polar disorder

Genotype	(11)	(12)	(13)	(22)	(23)	(33)	Total
Cases	271	2	0	371	0	104	748
Controls	306	0	0	333	0	86	725
Total	577	2	0	704	0	190	1473

Tintle et al. (2007) report a p-value of 0.061 from the χ_2^2 test ignoring inconsistently identified individuals and 0.064 from the test including inconsistencies. Using our software and assuming an additive mode of inheritance, the linear trend test just presented yields a p-value is 0.0230 ignoring inconsistents, 0.0241 including inconsistents using the method shown above and 0.0245 using a permutation test with 2000 permutations.

Cost-effectiveness of Duplicate Genotyping Using Previously Collected Data

Initially, we consider an instance of including previously collected quality control data in the test of association. In every case examined, the power of the LTT_d is higher when the duplicate genotype data is included as compared to when it is not. In other words, it is better to include the duplicate genotype data in subsequent tests of association then to ignore inconsistencies and treat the data as missing. This result is consistent with the results of Tintle et al. (2007) for the χ_2^2 test of association.

Evaluating the Cost-effectiveness of Collecting Duplicate Genotype Data

The most important case, however, is when $c>0$. That is, when we view the collection of duplicate genotype data as an *a priori* design decision, and thus must account for the cost of collecting the duplicates for a fraction, r, of the sample.

Given a fixed budget, in 49.2% of cases examined (see *Table 3*) where $c>0$, duplicate genotyping the entire sample ($r=1$) was found to be the most cost-effective design strategy (yields the highest power). In all remaining cases, $r=0$ provided the highest power. Thus, the optimal strategy is always "all or nothing."

In order to characterize situations where duplicate genotyping will be cost-effective, logistic regression models were used with all parameters predicting whether or not duplicate genotyping the entire sample was the most cost-effective design. Three parameters (δ_V, c and ε) had the strongest relationship with cost-effectiveness. Relative cost, c, had the strongest relationship (Wald χ^2=1424.2, *p<0.0001),* genotyping error rate ε also had a very strong relationship (Wald χ^2=1259.2, *p<0.0001)* and minor allele frequency (δ_V) was also strongly related (Wald χ^2=465.6, *p<0.0001).* As minor allele frequency (δ_V) declined, costs (c) declined, or genotyping error rate (ε) increased, duplicate genotyping the entire sample was more likely to be the optimal design decision.

Table 5 shows the percentage of cases examined in the computational study where duplicate genotyping the entire sample is the most effective design decision for different values of c (relative genotyping costs) and ε (genotyping error rate).

Table 5 Percent of cases where genotyping is cost-effective

	Genotyping error rate (ε)		
Relative cost (c)	0.001	0.01	0.03
0.001	100%	100%	100%
0.005	20%	100%	100%
0.01	0%	100%	100%
0.02	0%	76%	100%
0.05	0%	21%	87%
0.10	0%	0%	29%
0.50	0%	0%	0%

Table 5 demonstrates a general rule of thumb: duplicate genotyping the entire sample will always be cost-effective (regardless of δ_V) if $c \leq \varepsilon$. *Table* 5 also demonstrates that duplicate genotyping is sometimes cost-effective when $c>\varepsilon$. While details are not shown in the table, when δ_V is small, duplicate genotyping can be cost-effective even when $c>\varepsilon$.

Example Power Values

Table 6 provides power values for a specific example. Specifically, we present power under different values of δ_V, *ε, r* and *c* for a disease with a prevalence of 2.5%, disease allele frequency of 5%, equal number of cases and controls (*k*=1), and a SNP marker and disease allele that are in perfect linkage disequilibrium (ρ=1).

Table 6. Example of comparative power values

Marker Frequency (δ_V)	Genotyping error rate (ε)	Power if no duplicates (r=0)	Power if entire sample is duplicate genotyped (r=1)			
			c=0.001	c=0.01	c=0.1	c=0.5
Column I	Column II	Column III	Column IV	Column V	Column VI	Column VII
0.50	0.001	79.8	**79.8**	79.3	74.0	59.0
	0.01	78.0	**79.0**	**78.4**	73.1	58.0
	0.03	73.9	**76.9**	**76.3**	70.8	55.6
0.20	0.001	79.7	**79.8**	79.2	74.0	59.0
	0.01	77.3	**78.6**	**78.0**	72.7	57.6
	0.03	71.7	**75.8**	**75.2**	69.6	54.4
0.05	0.001	79.3	**79.6**	79.0	73.8	58.7
	0.01	73.1	**76.5**	**75.9**	70.4	55.1
	0.03	59.4	**69.4**	**68.7**	**62.9**	47.7

Bold *indicates that duplicate genotyping is cost-effective (power with r=1 has larger power than with r=0). Note that in all cases, if there were no misclassification errors and there was no duplicate genotyping (r=0), power would be 80%. The computations are based on disease allele frequency of 5%, additive mode of inheritance, a SNP marker and disease allele in perfect LD, 1.25x increased risk of disease if you have the disease allele, 2.5% of the population with the disease, an equal number of cases and controls (k=1) and α=0.0002.*

The sample size needed to yield 80% power was calculated assuming there was no genotyping error. *Column III* shows the power for that sample size, after taking the genotyping error into account. When the marker frequency is low and/or the genotyping error rate is larger *Column III* demonstrates that power can be significantly impacted by genotyping errors. *Columns IV-VII* then reduce the sample size to maintain the budget reflecting the additional cost of collecting duplicate genotype data on all samples at different genotype costs *(c)*. As *c* decreases, ε increases and δ_V decreases, *Columns IV-VIII* demonstrate that duplicate genotyping becomes more cost-effective. We note that in all cases duplicate genotyping does not meet or exceed the error-free power of 80% (detailed results not shown) however duplicate genotyping can successfully mediate some of the power loss due to genotyping errors.

Note that the power values in *Table 6* are from a specific example. Please use our software (*Results: Software*) to investigate power at values specific to your research situation while keeping in mind the rule of thumb presented in *Table 5*.

Recommendations for Use

In practice, duplicate genotyping should be considered when relative genotype to phenotype/sample acquisition costs do not exceed the expected SNP genotyping error rate. A more detailed treatment of practical considerations when using duplicate genotyping is provided in Tintle et al. (to appear). We summarize three main considerations here.

First, calculations provided in this manuscript consider only a single SNP. However, in practice, the decision to duplicate genotype will need to be made for an entire set of SNPs (e.g. all of the SNPs on a chip). In these cases using the same rule of thumb (duplicate genotype if $c \leq \varepsilon$) is appropriate where the ε used is the minimum error rate expected for any single SNP.

Second, if there is concern that some samples may be of low quality and, thus, have higher genotyping error rates than other samples (a violation of genotyping error assumption #2), the error rate, ε, used in the $c \leq \varepsilon$ rule of thumb should be the minimum expected ε for the high quality samples. Note, however, that we are still assuming non-differential errors in this case. Differential errors may increase the type I error rate, and are not considered in this manuscript.

Third, GWA studies are typically conducted in two-stages where all markers are genotyped on a sample of individuals, and then a subset of the markers is genotyped on a sample of additional individuals. When considering the use of duplicate genotyping in two-stage studies, the decision on the use of duplicate genotyping should be made separately at each stage since the relative cost of genotyping to phenotyping will be different at each stage.

Software

To facilitate the utilization of the methods discussed in this paper, we provide two companion pieces of software for this work. The first computes the LTT_d statistic and provides an asymptotic and permutation p-value. The second provides power computations for different genotyping costs, allele frequencies and error rates to assist in the duplicate genotyping design decision. Software is available at http://math.hope.edu/tintle/duplicate.html (source code written in *R*).

Conclusions

This work demonstrates how duplicate genotype data can be included in a linear trend test (*LTT)* of genetic association. Duplicate genotype data are included in the *LTT* through a weighting strategy and a subsequent adjustment of the variance of the *LTT* statistic yielding the LTT_d. We demonstrate via simulation that the asymptotic null and alternative distributions of the LTT_d statistic are obtained with reasonably small sample sizes in most cases. Both asymptotic and permutation test p-values are available in the free companion software.

We demonstrate that in the case of no duplicate genotyping costs (e.g. the data has already been collected) including the duplicate data in the LTT_d always increases statistical power. This confirms a similar result in Tintle et al. (2007).

We also consider the cost-effectiveness of designing a study to collect duplicate genotype data, and find that when the relative cost of genotyping to phenotype/sample acquisition costs *(c)* is less than or equal to the genotyping error rate *(ε)*, collecting duplicate genotype data on the entire sample is cost-effective. Further, we find that the optimal amount of duplicate genotyping, in these cases, will always involve duplicate genotyping the entire sample. In a two-stage GWA study for a complex disease, if a relatively small set of SNPs are being followed up at stage 2 and it will be costly to enroll more subjects, duplicate genotyping may be cost-effective since relative genotyping to phenotyping/acquisition costs *c* will be low.

Since the rule-of-thumb just described is conservative it is important to note that duplicate genotyping will be cost-effective in many situations when $c>\varepsilon$. This rule was provided to allow researchers to quickly assess the cost-effectiveness of duplicate genotyping on a large scale. It is quite likely that, even if $c>\varepsilon$, duplicate genotyping may provide moderate power gains for SNPs with low minor SNP allele frequency. Our software should be used to determine cost-effectiveness of duplicate genotyping for specific experimental conditions.

We assume that genotyping errors are independent from the first to second genotyping (genotyping error assumption #3) and that genotyping error rates are non-differential (genotyping error assumption #2). Future work is needed to extend results to consider differential genotyping errors when duplicate genotyping. Further reading on sources of genotyping error and their impact on analyses can be found in Bonin et al. (2004) and Gordon and Finch (2005). We also assume that duplicate genotyping is applied to a random subsample of size *nr*. Further work is necessary to explore optimizing the value of *r* depending upon phenotype or initial genotype classification.

When collected, duplicate genotype data should always be included in the subsequent test of association and in many realistic cases duplicate genotype data should be collected on the entire sample.

Appendix

Finding the Mean of U_d

$$\begin{aligned} E(U_d) &= E\left(\sum_i x_i\left(\frac{S}{N}t'_i - \frac{T}{N}s'_i\right)\right) \\ &= \sum_i x_i\left(\frac{STp_i^*}{N} - \frac{TSq_i^*}{N}\right) = \frac{ST}{N}\sum_i x_i\left(p_i^* - q_i^*\right) \end{aligned} \qquad (A1)$$

Finding the Variance of U_d

$$\begin{aligned} Var(U_d) &= Var\left(\sum_{i=1}^{3} x_i\left(\frac{S}{N}t'_i - \frac{T}{N}s'_i\right)\right) = \left(\frac{S}{N}\right)^2 Var\left(\sum_{i=1}^{3} x_i t'_i\right) + \left(\frac{T}{N}\right)^2 Var\left(\sum_{i=1}^{3} x_i s'_i\right) \\ &= \left(\frac{S}{N}\right)^2\left[\sum_{i=1}^{3}\left(x_i^2 Var(t'_i) + 2\sum_{\substack{j=2\\ i<j}}^{3} x_i x_j Cov(t'_i, t'_j)\right)\right] \\ &+ \left(\frac{T}{N}\right)^2\left[\sum_{i=1}^{3}\left(x_i^2 Var(s'_i) + 2\sum_{\substack{j=2\\ i<j}}^{3} x_i x_j Cov(s'_i, s'_j)\right)\right] \end{aligned} \qquad (A2)$$

Below we show how to find $Var(t'_1)$ and $Cov(t'_1, t'_2)$, other terms can be found similarly.

$$\begin{aligned} Var(t'_1) &= Var\left(t_1 + t_{11} + \frac{1}{2}\sum_{i\neq j}(t_{12} + t_{13})\right) = Var(t_1) + Var(t_{11}) + \frac{1}{4}Var(t_{12}) + \frac{1}{4}Var(t_{13}) \\ &+ Cov(t_{11}, t_{12}) + Cov(t_{11}, t_{13}) + \frac{1}{2}Cov(t_{12}, t_{13}) \\ &= T_S\, p_1^*\left(1 - p_1^*\right) + T_D\left(\begin{array}{l} p_{11}^*\left(1 - p_{11}^*\right) + \frac{1}{4}p_{12}^*\left(1 - p_{12}^*\right) + \frac{1}{4}p_{13}^*\left(1 - p_{13}^*\right) \\ - p_{11}^* p_{12}^* - p_{11}^* p_{13}^* - \frac{1}{2}p_{12}^* p_{13}^* \end{array}\right) \end{aligned}$$

$Cov(t'_1, t'_2) = E[t'_1\ t'_2] - E[t'_1] \cdot E[t'_2]$ where

$$E[t'_1] = E[t_1] + E[t_{11}] + \frac{1}{2}E[t_{12}] + \frac{1}{2}E[t_{13}] = NT_s p_1 + NT_d\left(p^*_{11} + \frac{1}{2}p^*_{12} + \frac{1}{2}p^*_{13}\right)$$

$$E[t'_2] = E[t_2] + E[t_{22}] + \frac{1}{2}E[t_{12}] + \frac{1}{2}E[t_{23}] = NT_s p_2 + NT_d\left(p^*_{22} + \frac{1}{2}p^*_{12} + \frac{1}{2}p^*_{23}\right)$$

$$E[t'_1\ t'_2] = E[t_1t_2] + E[t_1]E[t_{22}] + E[t_{11}]E[t_2] + E[t_{11}t_{22}] + \frac{1}{2}(E[t_1]E[t_{12}] + E[t_1]E[t_{23}] + E[t_{13}]E[t_2])$$
$$+\frac{1}{2}(E[t_{11}t_{12}] + E[t_{11}t_{23}] + E[t_{12}t_2] + E[t_{22}t_{12}] + E[t_{22}t_{13}]) + \frac{1}{4}(E[t_{12}t_{12}] + E[t_{12}t_{23}] + E[t_{12}t_{13}] + E[t_{13}t_{23}])$$

Where we make the following substitutions in $E[t'_1\ t'_2]$ as appropriate:

$E[t_i] = T_S p^*_i$, $E[t_{ij}] = T_D p^*_{ij}$ for $i = j$ or $i \neq j$, $E[t_i t_i] = Var(t_i) + E[t_i]^2$, $E[t_{ii}t_{ii}] = Var(t_{ii}) + E[t_{ii}]^2$, $E[t_i t_j] = Cov(t_i, t_j) + E[t_i]E[t_j]$ for $i \neq j$, and $E[t_{ij}t_{ik}] = Cov(t_{ij}t_{ik}) + E[t_{ij}]E[t_{ik}]$ for $j \neq k$.

Power of the LTT_d

Following the results of Zheng and Gastwirth (Zheng & Gastwirth 2006), asymptotic power for the LTT_d can be computed using the following formula

$$1 - \phi\left(\frac{z_{1-\alpha/2}\sigma_d - \mu_d}{\sigma_d}\right) + \phi\left(\frac{z_{\alpha/2}\sigma_d - \mu_d}{\sigma_d}\right) = 1 - \phi\left(z_{1-\alpha/2} - \frac{\mu_d}{\sigma_d}\right) + \phi\left(z_{\alpha/2} - \frac{\mu_d}{\sigma_d}\right) \quad (A3)$$

Where $\sigma_d = \sqrt{Var(U_d)}$ and $\mu_d = E(U_d)$.

References:

Ahn, K., Haynes, C., Kim, W., Fleur, R.S., Gordon, D. & Finch, S.J. (2007) The effects of SNP genotyping errors on the power of the Cochran-Armitage linear trend test for case/control association studies. *Annals of Human Genetics.* 71:Pt 2, 249-261.

Amos, C.I. (2007) Successful design and conduct of genome-wide association studies. *Human molecular genetics.* 16 Spec No. 2:R220-5.

Armitage, P. (1955) Tests for linear trends in proportions and frequencies. *Biometrics.* 11:375-386.

Bonin, A., Bellemain, E., Bronken Eidesen, P., Pompanon, F., Brochmann, C. & Taberlet, P. (2004) How to track and assess genotyping errors in population genetics studies. *Molecular ecology.* 13:11, 3261-3273.

Cochran, W.G. (1954) Some methods for strengthening the common chi-squared tests. *Biometrics.* 10:417-451.

Douglas, J.A., Skol, A.D. & Boehnke, M. (2002) Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *American Journal of Human Genetics.* 70:2, 487-495.

Freidlin, B., Zheng, G., Li, Z. & Gastwirth, J.L. (2002) Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Human heredity.* 53:3, 146-152.

Fridley, B.L., Turner, S.T., Chapman, A.B., Rodin, A.S., Boerwinkle, E. & Bailey, K.R. (2008) Reproducibiilty of genotypes as measured by the affymetrix GeneChip 100K Human Mapping Array set. *Computational Statistics and Data Analysis.* 52:5367-5374.

Gordon, D. & Finch, S.J. 2005, "Consequences of Error" in *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, eds. M.J. Dunn, L.B. Jorde, P.F.R. Little & S. Subramanian, Wiley, .

Gordon, D., Haynes, C., Yang, Y., Kramer, P.L. & Finch, S.J. (2007) Linear Trend Tests for Case-Control Genetic Association that Incorporate Random Phenotype and Genotype Misclassification Error. *Genetic epidemiology.* 31:853-870.

Gordon, D., Finch, S.J., Nothnagel, M. & Ott, J. (2002) Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Human heredity.* 54:1, 22-33.

Gordon, D. & Ott, J. (2001) Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. *Pacific Symposium on Biocomputing.Pacific Symposium on Biocomputing.* 18-29.

Gordon, D., Yang, Y., Haynes, C., Finch, S.J., Mendell, N.R., Brown, A.M. & Haroutunian, V. (2004) Increasing power for tests of genetic association in the presence of phenotype and/or genotype error by use of double-sampling. *Statistical applications in genetics and molecular biology.* 3:Article26.

Heid, I.M., Lamina, C., Kuchenhoff, H., Fischer, G., Klopp, N., Kolz, M., Grallert, H., Vollmert, C., Wagner, S., Huth, C., Muller, J., Muller, M., Hunt, S.C., Peters, A., Paulweber, B., Wichmann, H.E., Kronenberg, F. & Illig, T. (2008) Estimating the single nucleotide polymorphism genotype misclassification from routine double measurements in a large epidemiologic sample. *American Journal of Epidemiology.* 168:8, 878-889.

Kang, S.J., Gordon, D. & Finch, S.J. (2004) What SNP genotyping errors are most costly for genetic association studies? *Genetic epidemiology.* 26:2, 132-141.

Lai, R.Z., Zhang, H. & Yang, Y.N. (2007) Repeated measurement sampling in genetic association analysis with genotyping errors. *Genetic epidemiology.* 31:2, 143-153.

Pompanon, F., Bonin, A., Bellemain, E. & Taberlet, P. (2005) Genotyping errors: causes, consequences and solutions. *Nature reviews.Genetics.* 6:11, 847-859.

Rice, K.M. & Holmans, P. (2003) Allowing for genotyping error in analysis of unmatched case-control studies. *Annals of Human Genetics.* 67:Pt 2, 165-174.

Sasieni, P.D. (1997) From genotypes to genes: doubling the sample size. *Biometrics.* 53:4, 1253-1261.

Saunders, I.W., Brohede, J. & Hannan, G.N. (2007) Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference. *Genomics.* 90:3, 291-296.

Slager, S.L. & Schaid, D.J. (2001) Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. *Human heredity.* 52:3, 149-153.

Tintle, N.L., Gordon, D., Van Bruggen, D. & Finch, S.J. (to appear) The cost-effectiveness of duplicate genotyping for testing genetic association. *Ann.Hum.Genet.*

Tintle, N.L., Ahn, K., Mendell, N.R., Gordon, D. & Finch, S.J. (2005) Characteristics of replicated single-nucleotide polymorphism genotypes from COGA: Affymetrix and Center for Inherited Disease Research. *BMC genetics [computer file].* 6 Suppl 1:S154.

Tintle, N.L., Gordon, D., McMahon, F.J. & Finch, S.J. (2007) Using duplicate genotyped data in genetic analyses: testing association and estimating error rates. *Statistical applications in genetics and molecular biology.* 6:Article4.

Zheng, G., Freidlin, B., Li, Z. & Gastwirth, J.L. (2003) Choice of Scores in Trend Tests for Case-control studies of candidate-gene associations. *Biometrical journal.* 45:3, 335-348.

Zheng, G. & Gastwirth, J.L. (2006) On estimation of the variance in Cochran-Armitage trend tests for genetic association using case-control studies. *Statistics in medicine.* 25:18, 3150-3159.

Zuo, Y., Zou, G., Wang, J., Zhao, H. & Liang, H. (2008) Optimal two-stage design for case-control association analysis incorporating genotyping errors. *Annals of Human Genetics.* 72:Pt 3, 375-387.