Published in final edited form as:
Ann Hum Genet. 2009 May ; 73(Pt 3): 370–378. doi:10.1111/j.1469-1809.2009.00516.x.

The cost effectiveness of duplicate genotyping for testing genetic association

Nathan Tintle[1,*], **Derek Gordon**[2], **Dirk Van Bruggen**[1], and **Stephen Finch**[3]

[1] Hope College, Department of Mathematics, Holland, Michigan, USA

[2] Rutgers University, Department of Genetics, Piscataway, New Jersey, USA

[3] Stony Brook University, Department of Applied Mathematics and Statistics, Stony Brook, New York, USA

Summary

We consider a modification to the traditional genome wide association (GWA) study design: duplicate genotyping. Duplicate genotyping (re-genotyping some of the samples) has long been suggested for quality control reasons, however has not been evaluated for its statistical cost-effectiveness. We demonstrate that when genotyping error rates are at least *m*%, duplicate genotyping provides a cost-effective (more statistical power for the same price) design alternative when relative genotype to phenotype/sample acquisition costs are no more than *m*%. In addition to cost and error rate, duplicate genotyping is most cost-effective for SNPs with low minor allele frequency. In general, relative genotype to phenotype/sample acquisition costs will be low when following up a limited number of SNPs in the second stage of a two-stage GWA study design, and, thus, duplicate genotyping may be useful in these situations. In cases where many SNPs are being followed up at the second stage, duplicate genotyping only low-quality SNPs with low minor allele frequency may be cost-effective. We also find that in almost all cases where duplicate genotyping is cost-effective, the most cost-effective design strategy involves duplicate genotyping all samples. Free software is provided which evaluates the cost-effectiveness of duplicate genotyping based on user inputs.

Keywords

Genotype Error; Genome-wide Association Study; Re-genotype; Two-Stage; Power; Duplicate genotype

Introduction

Genome-wide association (GWA) studies are increasingly common. This popularity can be attributed to increased understanding of the human genome and readily accessible technology for measuring single nucleotide polymorphisms (SNPs) (Amos 2007). One remaining limitation of GWA studies is that they are very costly because of the large sample sizes needed to attain adequate statistical power to find the typically small effects of single genes on disease (Amos 2007). Thus, designs which can provide increased statistical power for the same cost as traditional designs are sorely needed in order to identify disease predisposing genetic variants more quickly and efficiently.

[*]Contact Author: Nathan Tintle, Ph.D., Assistant Professor of Mathematics, Hope College, Holland, MI 49423, tintle@hope.edu, 616-395-7272.

One such design, the two-stage design in its various forms (e.g. Skol et al. 2007, Zuo et al. 2008, Wang et al. 2006, Satagopan, Elston 2003), is an increasingly popular method of conducting GWA studies (Amos 2007). In a two-stage design, stage 1 evaluates many SNPs that are spread relatively uniformly across the genome. At stage 2 a larger sample is typically used, but analysis is only on certain regions of the genome or particular SNPs identified as interesting at stage 1. By focusing resources in regions of the genome that are of particular interest, two-stage designs yield a more powerful GWA study.

One promising area for further GWA studies design improvement (increased power) is in exploiting the limitations of SNP genotyping technology. Two such examples of technological exploitation providing increased power are double sampling, which consists of genotyping all individuals in the study and then using gene sequencing (viewed as a perfect method of genotyping) on a subset of the individuals (Gordon et al. 2004, Ji et al. 2005, Barral et al. 2006), and duplicate genotyping, which consists of re-running a subset of the samples (Tintle et al. 2007, Rice, Holmans 2003). Double sampling has been shown to be a cost-effective design when genotyping errors are present and increasingly so as the cost of genotyping declines (Ji et al. 2005, Levenstien, Ott & Gordon 2006). Including duplicate genotype data in the test of association has been shown to increase power when duplicate data has already been gathered (i.e. if you already have the duplicates, use them) (Tintle et al. 2007), but has not been evaluated for its *a priori* cost-effectiveness (i.e. should you collect duplicates?). Related to duplicate genotyping, Lai et al. (Lai, Zhang & Yang 2007) propose genotyping some individuals five times.

Genotyping errors are generally considered to have little analytic impact on results, due to low genotyping error rates at stage 1 of two-stage GWA studies (~0.1–0.2% or lower; Saunders, Brohede & Hannan 2007, Tintle et al. 2005). However, genotyping errors do not occur uniformly across all SNPs: some SNPs will have more errors than others (low quality SNPs). At stage 1 of a two-stage study, low quality SNPs are often considered to have little impact since standard chips use SNPs known to be typically measured with high quality, or if there is a low quality SNP, another high quality SNP known to be correlated (in high linkage disequilibrium) with the SNP of lower quality can be evaluated. In stage 2 analyses, however, a researcher may not merely be able to “move on” to the next SNP because the SNP of interest may not be in high linkage disequilibrium with other SNPs or the SNP of interest may be in a region with few known SNPs (for a visual depiction of a non-uniform SNP distribution in a region of interest see (Edwards et al. 2005)). In addition to stage 2 of two-stage studies, SNPs of low quality may be of interest if the study is focused on nonsynonymous SNPs that are directly related to increased disease risk (e.g. (Hampe et al. 2007)). Thus, we conclude that despite the overall decline in the impact of genotyping errors on analysis, there are some situations (i.e. nonsynonymous SNPs and stage 2 of two-stage designs) where genotyping errors may still be a significant issue due to the inherent inability to avoid low quality SNPs in these situations.

The effects of genotyping errors in testing genetic association are well known: differential genotyping error rates (cases and controls have different error models) increase Type I error rates (Moskvina et al. 2006). Non-differential genotyping error rates do not increase Type I error rates, but do increase type II error rates (Gordon, Ott 2001, Gordon et al. 2002, Ahn et al. 2007). Kang et al. (Kang et al. 2004, Kang, Gordon & Finch 2004) demonstrated that SNP genotyping errors impact statistical power the most when the SNP allele frequency associated with disease is low. Gordon and Finch provide reviews of much of the literature on this topic (Gordon, Finch 2005a, Gordon, Finch 2005b).

Recognizing that non-differential genotyping errors can reduce statistical power especially when SNP allele frequencies are low, and that at stage 2 of two-stage GWA studies and in

studies of non-synonymous SNPs genotyping errors can still be a measurable problem, we evaluate the cost-effectiveness of gathering duplicate genotype data. To do this we compare the statistical power of tests of association using duplicate genotyping data with designs that do not collect duplicate genotype data, but instead allocate resources to increase the sample size. As we will show, duplicate genotyping can be cost-effective depending on the cost of genotyping relative to sample acquisition costs (e.g. enrollment, phenotyping, sample preparation, etc.), the magnitude of the genotyping error rates, and the frequency of the SNP allele associated with disease.

Methods

Sampling Strategy

We consider a sampling strategy where a randomly selected fraction of the sample (r; $r \in [0,1]$) is genotyped twice, and the remaining fraction of the sample ($1-r$) is genotyped once. Here we assume that the same value of r is used for the entire sample. This may not be optimal for two-stage studies. For consideration of two-stage studies see *Results: Recommendations for use*.

Genotyping Error Assumptions

1. We assume that there is the possibility for genotyping errors. If $\varepsilon_{i,j}$ is the probability of an individual of genotype i being classified as genotype j, then we assume that $\varepsilon_{i,j}>0$ for at least one pair of i and j where $i \neq j$.
2. We assume that classification errors have the same probability for all individuals. This implies that genotyping errors are non-differential between cases and controls.
3. Finally, we assume independence of the classification errors. Specifically, this means that if $\varepsilon_{i,j}$ is the probability that an individual who is truly of genotype i is classified as genotype j, then regardless of what genotype the individual is actually classified to when first classified, that individual still has probability $\varepsilon_{i,j}$ of being classified to genotype j during the second classification. The assumption of independent errors for the duplicate genotyping strategy is crucial to its utility. If errors are not independent, then duplicate genotypings do not help to "clean up" the initial genotyping errors. Instead, the same individuals are incorrectly genotyped time after time. Tintle et al. (Tintle et al. 2005) suggest that the assumption of independent classification is met for SNP genotype data.

Power using duplicate genotyping

Tintle et al. (2007) discuss how to calculate the power of a test of association between phenotype and genotype that includes duplicate genotyped data and is asymptotically equivalent to the standard χ^2 test used for testing association. In this paper we consider the case of two phenotypes (e.g. case/control) and three genotypes (e.g. AA, AB and BB), by far the most popular testing scenario, though results are easily extendable to any number of genotypes/phenotypes.

Genetic Disease Model

To evaluate the cost-effectiveness of duplicate genotyping we assume that true genotype frequencies follow a standard genetic disease model (e.g. Amos 2007, Gordon et al. 2002, Ahn et al. 2007). This model is briefly outlined in the following sections.

Notation

p_2 = the frequency of allele *(2)* at the SNP marker. p_1= the frequency of allele *(1)* at the SNP marker. Thus, $p_1=1-p_2$. We assume that allele (*2*) is the allele associated with the disease risk allele *(D)* at the nearby disease locus.

p_D= the frequency of the disease risk allele (*D*) at the disease locus. p_+ = the frequency of the non-risk allele (+) at the disease locus. Thus, $p_+ = 1-p_D$.

h_{2D}=the frequency of the *2D* haplotype; that is, the frequency of having both the *2* allele at the SNP marker and the *D* allele at the disease locus. The other haplotypes are h_{1D}, h_{2+}, and h_{1+}. Thus, $h_{2D}+h_{1D}+h_{2+}+h_{1+}=1$.

D = the unstandardized measure of linkage disequilibrium between the SNP marker allele *(2)* associated with the disease risk allele *(D)* = $h_{2D}-p_2p_D$

r^2 = measure of the correlation between the SNP marker and the disease risk allele = $D^2/(p_2(1-p_2)p_D(1-p_D))$

$\rho = r^2/\max(r^2)$ = a measure of the correlation of allele *(2)* and the disease risk allele *(D)* as a fraction of their maximum possible correlation. As pointed out by Amos (2007), $\max(r^2)$ is not 1 unless $p_2=p_D$. We also note that for any values of p_2 and p_D, $\max(r^2)$ is attained when $D'=1$ where D' = the standardized measure of linkage disequilibrium where $D'= D/min(p_2(1-p_D),p_D(1-p_2))$

φ = the disease prevalence in the population

f_i = the penetrance of the disease given genotype *i* at the disease locus. Thus, f_{DD}= the probability someone who is *DD* at the disease locus (homozygote for the risk allele) has the disease, f_{+D}= the probability someone who is *+D* at the disease locus (heterozygote for the risk allele) has the disease, and f_{++}=the probability someone who is ++ at the disease locus (homozygous for the non-risk allele) has the disease.

γ =a general relative risk of disease parameter which is used to compute the genotype specific relative risks (γ_{DD} and γ_{+D}) in ways that are dependent upon the mode of inheritance of the disease (dominant, additive, recessive, multiplicative). Thus, γ_{DD} = the relative risk of disease (f_{DD}/f_{++}) for a participant with two copies of the risk allele. Similarly, γ_{+D} = the relative risk of disease (f_{+D}/f_{++}) for a participant with one copy of the risk allele. See Schaid and Sommer (1993).

Disease modes of inheritance

We consider four disease modes of inheritance. Specifically, dominant ($\gamma_{DD}=\gamma_{+D}=\gamma$), additive ($\gamma_{+D}=\gamma, \gamma_{DD}=2\gamma-1$), recessive ($\gamma_{DD}=\gamma$, $\gamma_{+D}=1$) and multiplicative ($\gamma_{+D}=\gamma, \gamma_{DD}=\gamma^2$).

Genotype error model parameters

$\varepsilon_{i,j}$ = the probability that an individual of genotype *i* is classified (called) as genotype *j*. Thus, when $i\neq j$, this is a genotype error probability.

The genotype error model of Douglas, Skol and Boehnke (2002) is considered in this paper. Specifically, we let $\varepsilon_{g1,g2} = \varepsilon_{g2,g1} = \varepsilon_{g3,g2} = \varepsilon_{g2,g3} = \varepsilon$ where we let g_1 represent the first genotype at the SNP marker (*11*), g_2 the second genotype at the SNP marker (*12*) and g_3 the third genotype at the SNP marker (*22*). Further we let $\varepsilon_{g1,g3} = \varepsilon_{g3,g1} = 0$. This model is line with the findings of Tintle et al. (2005).

Computational study

To evaluate the cost-effectiveness of duplicate genotyping, a computational study was conducted using the results of Tintle et al. (2007). The computational study was conducted as follows:

Step 1—For given values of p_2, p_D, ρ, φ, γ and the disease mode of inheritance the true genotype frequencies were computed. (see the Appendix for details)

Step 2—For given values of k (the ratio of controls to cases), α (in this study always set to 1/300,000 to approximate an α that would be used to adjust for multiple testing e.g. Skol et al. 2007), and the power of the test if there were no genotyping errors and no duplicate genotyping, the equations of Tintle et al. (2007) can be used to find the number of cases and controls needed to attain the given power. We note that other computational studies with different α's have yielded similar results to what is shown in the upcoming sections though those results are not presented.

Step 3—Let c be the relative cost of genotyping to phenotyping and acquisition costs. Then for any r (the duplicate genotyping fraction) and ε, the power of the test of association using duplicate genotyping can be computed. To ensure appropriate comparisons we let $B=(1+c)n$ where B= total budget and n=sample size if no duplicate genotyping. This budget equation assumes that all individuals in the study will have the same relative genotyping to phenotyping costs, c. This is typically not the case for a two-stage design (relative costs for the stage 1 sample are different than relative costs in the stage 2 sample). For discussion of the two-stage design see *Results: Recommendations for use*. The budget if there were no duplicate genotyping is then "spent" for a duplicate genotyping design. We use the following equation to find the value of n_r which corresponds to the same costs (B) as the no duplicate genotyping design: $B=(1+c)n_r(1-r)+(1+2c)n_r r$. In the computational study we examined values of the variables as shown in Table 1.

We examine all possible combinations of values, thus, a total of 1,037,232 settings were evaluated (7×7×2×3×2×7×7×3×4×3). Further, for each setting of the computational study, power was computed for multiple values of r, specifically r=0, 0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90 and 1.00.

Software

Software that performs the power calculation is available for download from the author's website (http://math.hope.edu/tintle/duplicate.html) (Note: source code written in *R*).

Results

Optimally designed duplicate genotype study

Given a fixed budget (B) for sampling (acquisition and classification costs) spent in a variety of different ways (different values of r; $0 \le r \le 1$), and holding all other variables constant we searched for the value of r which maximized the power of the test of association. Overall, in 47.8% (495,795/1,037,232) of the situations examined, duplicate genotyping the entire sample $(r=1)$ was optimal since it yielded the highest power. In most other cases (52.16%; 540,996/1,037,232) the optimal proportion for duplicate genotyping was 0. In the few remaining cases (0.04%; 441/1,037,232) where the optimal design was to duplicate genotype a fraction of the samples ($0<r<1$), the design would be nearly optimized (within 0.038% of optimal) by either genotyping everyone exactly twice (complete duplicate genotyping) or everyone exactly once (no duplicate genotyping). Thus, we conclude that, from a statistical power perspective, the optimal study design accounting for genotype errors

using duplicate genotyping is an all or nothing proposition: one should either duplicate genotype everyone or no one.

The cost-effectiveness of duplicate genotyping

To examine which parameters were associated with complete duplicate genotyping *(r=1)* being the optimal design, we used logistic regression predicting whether or not *r=1* was optimal. We fit separate logistic regression models for each of the four different disease inheritance models, using p_2, p_D, ρ, disease prevalence (φ), genotyping error rate (ε), relative cost of genotyping to sample acquisition and phenotyping (c), risk of disease (γ), ratio of controls to cases (k) and power if there were no genotyping errors (80% or 95%) as explanatory variables in the models.

We say that duplicate genotyping is cost-effective if the optimal value of r ($r \in [0,1]$) is 1. The three strongest associations with the cost-effectiveness of duplicate genotyping were the genotyping error rate (ε), the relative cost of genotyping (c) and the allele frequency of the SNP (p_2). In all four disease inheritance models, as the error rate increased, cost of genotyping decreased, or p_2 moved farther from 0.5, duplicate genotyping was more likely to be cost-effective. Some of the other variables had substantially weaker, but sometimes significant, effects in some of the inheritance models (results not shown).

Since genotyping costs and error rates are the most related to the cost-effectiveness of duplicate genotyping, Table 2 shows the percentage of cases we examined in which duplicate genotyping was cost-effective for different combinations of genotyping error rate and relative genotyping cost. Thus, regardless of other model parameters, for genotyping error rates of at least 0.1%, duplicate genotyping is cost effective as long as the relative cost of genotyping to phenotyping/sample acquisition is no more than 0.1%. Similarly, for genotyping error rates of at least 0.5%, 1.0% and 5%, duplicate genotyping is cost-effective for relative genotyping costs no larger than 0.5%, 1.0% and 5%, respectively. In general, duplicate genotyping is cost-effective as long as $c \leq \varepsilon$.

Percentage of power loss reclaimed through duplicate genotyping

It is well documented that power loss from genotyping errors can be substantial (Gordon et al. 2002). Since in many cases it is most cost-effective to duplicate genotype the entire sample, it is important to note that the power gain from duplicate genotyping never exceeds the power of the error-free study (results not shown). However, it is of interest to know how much of the power loss due to genotyping errors can be “reclaimed” via duplicate genotyping.

Across the many situations where duplicate genotyping was found to be cost-effective, duplicate genotyping reclaims up to 57% of the power loss due to genotyping errors where the reclamation percentage is computed as

$$\text{Reclamation Percentage} = \frac{Power(r=1 \text{ with errors}) - Power(r=0 \text{ with errors})}{Power(r=0 \text{ and no errors}) - Power(r=0 \text{ with errors})}$$

Multiple linear regression models predicting reclamation percentage yielded similar results to the logistic regression models predicting whether or not duplicate genotyping was cost-effective. Specifically, as the genotyping error rates increased, relative genotyping costs decreased and SNP allele frequency (p_2) decreased, reclamation percentage increased.

Examples of power gain from duplicate genotyping

To illustrate actual power gain from duplicate genotyping we present some example power values. Consider a situation where we want to design a study with 80% power to detect a risk of 1.50 for a dominant disease with prevalence 5% using α=1/300,000 and an equal number of cases and controls (k=1). Further, assume that the SNP marker allele and the disease allele are in perfect linkage disequilibrium (they attain their maximum r^2; $\rho=1$), and that the disease allele frequency is 5%. Results in Table 3 consider a variety of genotyping error rates, relative costs and SNP marker allele frequencies.

Table 3 shows the power gain/loss from duplicate genotyping the entire sample. The initial sample sizes (n) used in this table (if there were no duplicate genotyping; $r=0$) were calculated to yield 80% power if there were no errors. The $r=0$ column in the table presents powers less than 80%, because there are genotyping errors ($\varepsilon>0$), and so using only n samples yields power less than 80%. A substantial reduction in power when the SNP marker frequency is low and/or the genotyping error rate is large is evident from the $r=0$ column. The last four columns of the table have reduced the total sample size (n) to account for the costs of duplicate genotyping the entire sample ($r=1$). In these columns, power values are bolded when they are larger than the power if $r=0$, reflecting situations where duplicate genotyping the entire sample ($r=1$) yields more power than not duplicate genotyping ($r=0$) at the same total design cost.

For example, when the SNP allele frequency (p_2) is 0.05 and the genotyping error rate (ε) is 0.5%, a study that was designed to yield 80% power (necessary n=7,058) if there were no genotyping errors only has 76.00% power because of the genotyping errors. If the cost of phenotyping and sample acquisition is $1,000 per subject and the cost of genotyping a set of SNPs of interest is $10 per person (c=0.01), then the total budget needed is $7,128,580 at a cost of $1010 per person. If money is re-allocated from sample acquisition/phenotyping (i.e. reduce the sample size to n=6,989) and put towards duplicate genotyping everyone in the sample ($r=1$), effectively costing $1020 per person instead of $1010, the power is 77.18%.

The table demonstrates the general rule of thumb (duplicate genotyping is cost-effective if $c\leq\varepsilon$) presented in Table 2. For example, duplicate genotyping (r=1) provides a higher power than single genotyping when c=0.001 for error rates of at least 0.1% for all marker frequencies. While actual power gains can sometimes be small, when SNP allele frequencies are small and costs are low, duplicate genotyping can provide moderate gains in power, even for relatively low genotyping error rates

Because of the dynamic genotyping environment, costs continue to change quickly. With this in mind, software, written in *R*, is available (http://math.hope.edu/tintle/duplicate.html) to conduct power calculations for the most current cost and genetic model information.

Recommendations for use

SNPs can suffer from significant power loss if they have measurable amounts of genotyping errors, especially at low allele frequencies. It is exactly these SNPs that will benefit the most from duplicate genotyping when relative genotype to phenotype costs (c) are low.

In practice, it is often the case that the decision to duplicate genotype cannot be made on a SNP-by-SNP basis. Instead, the decision must be made for a set of SNPs (e.g. all of the SNPs on a chip) that have many different error rates and many different allele frequencies. Recall the rule-of-thumb presented earlier: duplicate genotyping is cost-effective, regardless of allele frequency, whenever $c\leq\varepsilon$. When deciding whether or not to duplicate genotype a set of SNPs, if $c\leq\varepsilon$, where ε is the lowest expected ε for any of the SNPs in the set, then duplicate genotyping will be cost-effective for all of the SNPs under study. Cost-

effectiveness, however, does not mean that all of the SNPs will benefit from substantial power gains. On the contrary, the SNPs with larger minor allele frequencies (0.10 and higher) and/or lower error rates will have minimal power gain from duplicate genotyping. However, SNPs with lower minor allele frequencies and or higher error rates stand to benefit from measurable power gains. Using the rule-of-thumb, the design decision to duplicate genotype will have benefited many SNPs in a small way, and some SNPs in a moderate to large way.

Use of duplicate genotyping in two-stage designs requires some special considerations since the relative cost of genotyping will likely be different at stage 1 then at stage 2 and, up to this point, we have assumed that everyone in the sample has the same genotyping costs. To make the decision about whether or not to duplicate genotype in one, both or neither stages, we note that the most cost-effective sample design for duplicate genotyping is only affected by the relative cost of genotyping (c), but not the sample size/budget. Thus, one should independently decide if duplicate genotyping is cost-effective in the first stage (in today's marketplace likely not) and the second stage (in today's marketplace, more likely). For example, we propose a two part optimization process. First, the two-stage design is optimized to find the number of subjects and markers to analyze in each stage based on genotyping, sampling and phenotyping costs and SNP specific parameters (Muller, Pahl & Schafer 2007). Then, each stage can be evaluated separately for the cost-effectiveness of using duplicate genotyping using the relative genotyping costs and expected error rates of SNPs in that stage. If duplicate genotyping is cost-effective for either stage, then the number of individuals sampled at that stage is reduced to maintain the budget allocated to that stage. If duplicate genotyping is not cost-effective for either stage, then the number of individuals sampled at that stage is kept the same.

Instead of duplicate genotyping all SNPs under study, targeting only those SNPs with larger error rates or lower allele frequencies for duplicate genotyping is also a reasonable approach, as it is these SNPs that yield the largest increases in power. However, care must be taken to evaluate the impact of such a design decision on the relative cost, c. For example, duplicate genotyping only certain SNPs likely means the creation of a custom SNP chip that may be costly. This decision, as with all design decisions considered in this paper, is highly dependent on the relative cost, c, which is a highly volatile parameter in today's market.

While genotyping error assumption #2 states that error rates will be the same for all individuals, this may not always be true in practice (e.g. cases where some samples are of low quality). To ensure the proper global design decision, the error rate (ε) used should be the minimum expected error rate for the high-quality DNA samples. Of course, duplicate genotyping will be the most cost-effective for the most error prone samples, and so, once samples have been collected, error prone samples may benefit from re-genotyping as long as the sample does not also violate genotyping assumption #3.

Discussion

In this paper we have evaluated the cost-effectiveness of duplicate genotyping using the alternative hypothesis distribution for the duplicate genotyping test statistic as presented by Tintle et al. (2007). In order to evaluate the cost-effectiveness of duplicate genotyping we compared different allocations of resources (budget) to sample acquisition/phenotyping and genotyping and found the resulting power. The sample design which yielded the highest power for a fixed budget was deemed optimal. We found that in many cases duplicate genotyping all samples was the optimal design. As a general rule-of-thumb duplicate genotyping the entire sample is cost-effective as long as the relative cost of genotyping to

phenotyping is less than the genotyping error rate. The power gains from duplicate genotyping can become moderate for SNPs with small minor allele frequencies, even when genotyping error rates are low (≥0.5%).

There are many sample designs possible for conducting genotype-phenotype association studies. We have discussed some popular designs in the *Recommendations for Use* section. In general, when the cost of duplicate genotyping is inexpensive, when the cost of acquiring and phenotyping additional samples is expensive (for example, complex or rare phenotypes), when the genotyping error rate is high (for example, when the SNPs of interest are of low quality but cannot be avoided), or when the minor allele frequency is low duplicate genotyping may be a cost effective sample design. The free software provided can assist in the design decision.

Our findings are in line with previous research in this area. Specifically, Kang et al. (2004) showed that genotyping errors reduce power the most when allele frequency is small. Similarly, we found that duplicate genotyping regained the most power loss in situations where allele frequency was low. Further, previous research by Ji et al. (2005) and Levenstein et al. (2006) showed that double sampling (genotyping the entire sample and sequencing a subsample; see the *Introduction*) was the most cost-effective when genotyping was inexpensive. Similarly, we have shown that duplicate genotyping is the most cost-effective when genotyping is inexpensive relative to phenotyping and sample acquisition. Further work is needed to compare duplicate and double sampling.

Duplicate genotyping can substantially mediate the negative impact of genotyping errors on statistical power. Overall, when genotyping is cheap, duplicate genotyping will provide the biggest advantages when error rates are large and allele frequencies are small. With this in mind, duplicate genotyping may be most beneficial in classifying individuals using new technology where error rates are higher. In particular, copy number variation technology, which has small allele frequencies and may have moderate rates of classification error (Kim et al. 2008), may provide another realistic setting with which to use duplicate classification, though further work is necessary to implement duplicate genotyping in that context.

The market conditions for genotyping continue to rapidly change. New technologies and refinements to old technologies are rapidly appearing, genotyping costs are continuing to drop overall, and genotyping errors remain a small, but consistent problem. In this dynamic environment and with GWA studies working to increase power in any way possible, considerations should be given to alternative sample designs if they can be shown to increase power at a fixed cost. In this paper we have shown that for reasonable cost and error settings, duplicate genotyping provides a cost-effective design, with measurable power gains when SNPs of interest have small allele frequencies.

Acknowledgments

We thank Airat Bekmetjev and two anonymous reviewers for helpful feedback on this manuscript. 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, Dirk Van Bruggen received partial support from a computational science and modeling scholar award from the Hope College Howard Hughes Medical Institute program, a fellowship from the Michigan Space Grant Consortium and support from the Tanis Fund for Statistics Research.

References

Ahn K, Haynes C, Kim W, Fleur RS, Gordon D, Finch SJ. 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. 2007; 71(Pt 2):249–261. [PubMed: 17096677]

Amos CI. Successful design and conduct of genome-wide association studies. Human molecular genetics. 2007; 16(Spec No 2):R220–5. [PubMed: 17597095]

Barral S, Haynes C, Stone M, Gordon D. LRTae: improving statistical power for genetic association with case/control data when phenotype and/or genotype misclassification errors are present. BMC genetics. 2006; 7:24. [PubMed: 16689984]

Douglas JA, Skol AD, Boehnke M. Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. American Journal of Human Genetics. 2002; 70(2):487–495. [PubMed: 11791214]

Edwards AO, Ritter R 3rd, Abel KJ, Manning A, Panhuysen C, Farrer LA. Complement factor H polymorphism and age-related macular degeneration. Science (New York, NY). 2005; 308(5720): 421–424.

Gordon, D.; Finch, SJ. Consequences of Error. In: Dunn, MJ.; Jorde, LB.; Little, PFR.; Subramanian, S., editors. Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. Wiley; 2005a.

Gordon D, Finch SJ. Factors affecting statistical power in the detection of genetic association. The Journal of clinical investigation. 2005b; 115(6):1408–1418. [PubMed: 15931375]

Gordon D, Finch SJ, Nothnagel M, Ott J. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. Human heredity. 2002; 54(1):22–33. [PubMed: 12446984]

Gordon D, Ott J. Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. Pacific Symposium on Biocomputing. 2001:18–29. [PubMed: 11262939]

Gordon D, Yang Y, Haynes C, Finch SJ, Mendell NR, Brown AM, Haroutunian V. 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. 2004; 3 Article26.

Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, Albrecht M, Mayr G, De La Vega FM, Briggs J, Gunther S, Prescott NJ, Onnie CM, Hasler R, Sipos B, Folsch UR, Lengauer T, Platzer M, Mathew CG, Krawczak M, Schreiber S. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nature genetics. 2007; 39(2):207–211. [PubMed: 17200669]

Ji F, Yang Y, Haynes C, Finch SJ, Gordon D. Computing asymptotic power and sample size for case-control genetic association studies in the presence of phenotype and/or genotype misclassification errors. Statistical applications in genetics and molecular biology. 2005; 4 Article37.

Kang SJ, Finch SJ, Haynes C, Gordon D. Quantifying the percent increase in minimum sample size for SNP genotyping errors in genetic model-based association studies. Human heredity. 2004; 58(3–4):139–144. [PubMed: 15812170]

Kang SJ, Gordon D, Finch SJ. What SNP genotyping errors are most costly for genetic association studies? Genetic epidemiology. 2004; 26(2):132–141. [PubMed: 14748013]

Kim W, Gordon D, Sebat J, Ye KQ, Finch SJ. Computing power and sample size for case-control association studies with copy number polymorphism: application of mixture-based likelihood ratio test. PLoS ONE. 2008; 3(10):e3475. [PubMed: 18941524]

Lai RZ, Zhang H, Yang YN. Repeated measurement sampling in genetic association analysis with genotyping errors. Genetic epidemiology. 2007; 31(2):143–153. [PubMed: 17187401]

Levenstien MA, Ott J, Gordon D. Are molecular haplotypes worth the time and expense? A cost-effective method for applying molecular haplotypes. PLoS genetics. 2006; 2(8):e127. [PubMed: 16933998]

Moskvina V, Craddock N, Holmans P, Owen MJ, O'Donovan MC. Effects of differential genotyping error rate on the type I error probability of case-control studies. Human heredity. 2006; 61(1):55–64. [PubMed: 16612103]

Muller H, Pahl R, Schafer H. Including sampling and phenotyping costs into the optimization of two stage designs for genomewide association studies. Genetic epidemiology. 2007; 31:844–852. [PubMed: 17549751]

Rice KM, Holmans P. Allowing for genotyping error in analysis of unmatched case-control studies. Annals of Human Genetics. 2003; 67(Pt 2):165–174. [PubMed: 12675691]

Satagopan JM, Elston RC. Optimal two-stage genotyping in population-based association studies. Genetic epidemiology. 2003; 25(2):149–157. [PubMed: 12916023]

Saunders IW, Brohede J, Hannan GN. Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference. Genomics. 2007; 90(3):291–296. [PubMed: 17587543]

Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate-gene association studies. 1993

Skol AD, Scott LJ, Abecasis GR, Boehnke M. Optimal designs for two-stage genome-wide association studies. Genetic epidemiology. 2007; 31(7):776–788. [PubMed: 17549752]

Tintle NL, Ahn K, Mendell NR, Gordon D, Finch SJ. Characteristics of replicated single-nucleotide polymorphism genotypes from COGA: Affymetrix and Center for Inherited Disease Research. BMC genetics [computer file]. 2005; 6(Suppl 1):S154.

Tintle NL, Gordon D, McMahon FJ, Finch SJ. Using duplicate genotyped data in genetic analyses: testing association and estimating error rates. Statistical applications in genetics and molecular biology. 2007; 6 Article 4.

Wang H, Thomas DC, Pe'er I, Stram DO. Optimal two-stage genotyping designs for genome-wide association scans. Genetic epidemiology. 2006; 30(4):356–368. [PubMed: 16607626]

Zuo Y, Zou G, Wang J, Zhao H, Liang H. Optimal two-stage design for case-control association analysis incorporating genotyping errors. Annals of Human Genetics. 2008; 72(Pt 3):375–387. [PubMed: 18215207]

Appendix

Computing the true genotype frequencies

This section describes how to use the values of p_2, p_D, ρ, φ, γ, the disease mode of inheritance and ε to find the true genotype frequencies in the cases and controls. Recall that ρ is a number between 0 and 1 which is multiplied by the maximum r^2 in order to find the actual r^2 between the SNP allele and the disease allele (when ρ=1 the marker and disease alleles are in perfect linkage disequilibrium)..

Step 1

Use p_2, p_D, and ρ, to find the haplotype frequencies h_{2D}, h_{1D}, h_{2+}, and h_{1+} as shown below.

Step 1a—As is shown in Amos (2007), if r^2 is maximized than the following is true:

$$p_{2D} = \min(p_D(1 - p_2), p_2(1 - p_D)) + p_D * p_2.$$

Thus, $$\max(r^2) = \frac{(p_{2d} - p_D p_2)^2}{(p_2[1 - p_2]p_D[1 - p_D])}.$$

Step 1b—To find the observed r^2 multiply max(r^2) times ρ.

Step 1c—Find D using the observed r^2 from *Step 1b*.

$$D = \sqrt{r^2 p_2(1-p_2)p_D(1-p_D)}$$

Step 1d—Find the haplotype probabilities as follows:

$$h_{1+} = (1-P_D)(1-p_2) + D$$
$$h_{2+} = (1-P_D)p_2 - D$$
$$h_{1d} = p_D(1-p_2) - D$$
$$h_{2d} = p_D p_2 + D$$

Step 2

Use the disease mode of inheritance, φ, and γ, to find the disease penetrances.

If we assume that the disease locus is in Hardy-Weinberg Equilibrium then it can be shown (see, for example, Gordon et al. 2002) that

$$f_{++} = \frac{\varphi}{(1-p_D)^2 + 2p_D(1-p_D)\gamma_{+D} + p_D^2\gamma_{DD}}$$
$$f_{+D} = f_{++}\gamma_{+D}$$
$$f_{DD} = f_{++}\gamma_{DD}$$

Step 3

The disease penetrances from *Step 2* can be used along with haplotype frequencies from *Step 1* to find the true genotype distributions in both the cases (with disease; denoted *A*) and controls (without disease; denoted *U*). As noted earlier, we let g_1 stand for the first genotype at the SNP marker (namely, *11*), g_2 for the second genotype (*12*) and g_3 for the third genotype (*22*).

Let ${}_C p_g$ = the probability that an individual with disease status *C (=A,U)* is of genotype *g* (g_1, g_2,g_3). It is shown (Ahn et al. 2006) that

$${}_A p_{g_1} = \frac{1}{\varphi}\left[(h_{+1})^2 f_{++} + 2(h_{+1}h_{D1})f_{+D} + (h_{D1})^2 f_{DD}\right]$$
$${}_A p_{g_2} = \frac{2}{\varphi}\left[(h_{+1}h_{+2})\, f_{++} + (h_{+1}h_{D2} + h_{+2}h_{D1})f_{+D} + (h_{D1}h_{D2})f_{DD}\right]$$
$${}_A p_{g_3} = \frac{1}{\varphi}\left[(h_{+2})^2 f_{++} + 2(h_{+2}h_{D2})f_{+D} + (h_{D2})^2 f_{DD}\right]$$
$${}_U p_{g_1} = \frac{1}{1-\varphi}\left[(h_{+1})^2(1-f_{++}) + 2(h_{+1}h_{D1})(1-f_{+D}) + (h_{D1})^2(1-f_{DD})\right]$$
$${}_U p_{g_2} = \frac{2}{1-\varphi}\left[(h_{+1}h_{+2})\,(1-f_{++}) + (h_{+1}h_{D2} + h_{+2}h_{D1})(1-f_{+D}) + (h_{D1}h_{D2})(1-f_{DD})\right]$$
$${}_U p_{g_3} = \frac{1}{1-\varphi}\left[(h_{+2})^2(1-f_{++}) + 2(h_{+2}h_{D2})(1-f_{+D}) + (h_{D2})^2(1-f_{DD})\right]$$

Finding the observed genotype frequencies

If we know the true genotype frequencies, we can apply the genotype error model and find the observed (in the presence of genotyping errors) genotypes as shown below.

$$
\begin{aligned}
{}_A p^*_{g_1} &= {}_A p_{g_1}(1-\varepsilon) + {}_A p_{g_2}\varepsilon \\
{}_A p^*_{g_2} &= {}_A p_{g_1}\varepsilon + {}_A p_{g_2}(1-\varepsilon) + {}_A p_{g_3}\varepsilon \\
{}_A p^*_{g_3} &= {}_A p_{g_2}\varepsilon + {}_A p_{g_3}(1-\varepsilon)
\end{aligned}
$$

Further, we can find the observed genotype frequencies after duplicate genotyping recalling genotyping error assumption #3, that genotyping errors are independent. Here, ${}_C p^*_{g\,A,gB}$ represents the probability that an individual of disease status *C (=A,U)* is observed in genotype *A* one time and genotype *B* one time when duplicate genotyped.

$$
\begin{aligned}
{}_A p^*_{g_1,g_1} &= {}_A p_{g_1}(1-\varepsilon)^2 + {}_A p_{g_2}\varepsilon^2 \\
{}_A p^*_{g_1,g_2} &= {}_A p_{g_1}(1-\varepsilon)\varepsilon + {}_A p_{g_2}(1-\varepsilon)\varepsilon \\
{}_A p^*_{g_2,g_2} &= {}_A p_{g_1}\varepsilon^2 + {}_A p_{g_2}(1-\varepsilon)^2 + {}_A p_{g_3}\varepsilon^2 \\
{}_A p^*_{g_1,g_3} &= {}_A p_{g_2}\varepsilon^2 \\
{}_A p^*_{g_2,g_3} &= {}_A p_{g_2}(1-\varepsilon)\varepsilon + {}_A p_{g_3}(1-\varepsilon)\varepsilon \\
{}_A p^*_{g_1,g_1} &= {}_A p_{g_3}(1-\varepsilon)^2 + {}_A p_{g_2}\varepsilon^2
\end{aligned}
$$

Table 1

Variable values for the computational study

Variable	Values considered
p_2	0.01, 0.05, 0.10, 0.30, 0.50, 0.70, 0.90
p_d	0.01, 0.05, 0.10, 0.30, 0.50, 0.70, 0.90
ρ	0.8, 1
φ	0.01, 0.05, 0.10
Power if ε=0 (no errors)	0.80, 0.95
ε	0.0005, 0.001, 0.002, 0.005, 0.01, 0.03, 0.05
c	0.001, 0.005, 0.01, 0.02, 0.05, 0.10, 0.50
γ	1.25, 1.50, 2.00
Disease mode of inheritance	Dominant, Additive, Recessive, Multiplicative
k	1, 0.5, 2

Table 2

Percentage of cases examined where duplicate genotyping is cost-effective

Relative cost of genotyping (c)	Genotyping error rate (ε)						
	0.0005	**0.001**	**0.002**	**0.005**	**0.01**	**0.03**	**0.05**
0.001	62.5	100.0	100.0	100.0	100.0	100.0	100.0
0.005	16.3	29.1	60.2	100.0	100.0	100.0	100.0
0.01	13.1	16.3	29.2	63.3	100.0	100.0	100.0
0.02	1.3	13.1	16.3	32.4	63.6	100.0	100.0
0.05	0.0	0.9	5.5	16.3	29.3	90.4	100.0
0.10	0.0	0.0	0.9	13.2	16.3	49.9	71.1
0.50	0.0	0.0	0.0	0.0	0.5	15.0	16.4

Table 3

An example of power comparisons

			Power if complete duplicate genotyping (r=1)			
Marker Frequency (p_2)	**Genotyping error rate (ε)**	**Power if no duplicate genotyping (r=0)**	**c=0.001**	**c=0.01**	**c=0.1**	**c=0.5**
0.50	0.0005	79.88%	79.86%	79.14%	72.44%	52.94%
	0.001	79.76%	**79.80%**	79.08%	72.37%	52.86%
	0.002	79.51%	**79.68%**	78.96%	72.23%	52.71%
	0.005	78.78%	**79.31%**	78.58%	71.82%	52.27%
	0.01	77.52%	**78.68%**	**77.95%**	71.11%	51.52%
	0.03	72.16%	**76.03%**	**75.26%**	68.16%	48.48%
	0.05	66.38%	**73.16%**	**72.35%**	65.03%	45.39%
0.30	0.0005	79.87%	79.85%	79.14%	72.43%	52.92%
	0.001	79.74%	**79.79%**	79.07%	72.36%	52.84%
	0.002	79.47%	**79.66%**	78.93%	72.21%	52.68%
	0.005	78.66%	**79.25%**	78.53%	71.75%	52.19%
	0.01	77.28%	**78.57%**	**77.83%**	70.98%	51.38%
	0.03	71.41%	**75.67%**	**74.89%**	67.76%	48.07%
	0.05	65.08%	**72.50%**	**71.69%**	64.33%	44.71%
0.10	0.0005	79.76%	**79.80%**	79.08%	72.37%	52.83%
	0.001	79.53%	**79.68%**	78.96%	72.23%	52.69%
	0.002	79.05%	**79.44%**	78.72%	71.96%	52.40%
	0.005	77.58%	**78.72%**	**77.98%**	71.14%	51.53%
	0.01	75.07%	**77.47%**	**76.72%**	69.75%	50.07%
	0.03	64.32%	**72.11%**	**71.29%**	63.90%	44.28%
	0.05	53.27%	**66.20%**	**65.34%**	**57.73%**	38.62%
0.05	0.0005	79.61%	**79.73%**	79.00%	72.27%	52.70%
	0.001	79.22%	**79.53%**	78.81%	72.05%	52.47%
	0.002	78.43%	**79.14%**	78.41%	71.60%	51.99%
	0.005	76.00%	**77.93%**	**77.18%**	70.25%	50.57%
	0.01	71.80%	**75.85%**	**75.08%**	67.95%	48.22%

			Power if complete duplicate genotyping (r=1)			
Marker Frequency (p_2)	**Genotyping error rate (ε)**	**Power if no duplicate genotyping (r=0)**	**c=0.001**	**c=0.01**	**c=0.1**	**c=0.5**
	0.03	54.46%	**66.82%**	**65.97%**	**58.36%**	39.16%
	0.05	38.95%	**57.21%**	**56.33%**	**48.75%**	31.06%
0.01	0.0005	78.27%	**79.06%**	**78.33%**	71.53%	51.96%
	0.001	76.51%	**78.18%**	**77.44%**	70.55%	50.92%
	0.002	72.92%	**76.41%**	**75.64%**	68.57%	48.89%
	0.005	62.06%	**70.95%**	**70.12%**	**62.67%**	43.14%
	0.01	45.68%	**61.74%**	**60.87%**	**53.23%**	34.75%
	0.03	12.74%	**32.23%**	**31.53%**	**25.89%**	**14.72%**
	0.05	4.68%	**17.26%**	**16.82%**	**13.38%**	**7.14%**

*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%, dominant mode of inheritance, a SNP marker and disease allele in perfect LD, 1.5x increased risk of disease if you have the disease allele, 5% of the population with the disease, an equal number of cases and controls (k=1) and α=1/300,000.