A Comparison of Approaches to Account for Uncertainty in Analysis of Imputed Genotypes

Jin Zheng,[1] Yun Li,[1] Gonçalo R. Abecasis,[1] and Paul Scheet[1,2]*

[1]*Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan*
[2]*Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, Texas*

The availability of extensively genotyped reference samples, such as "The HapMap" and 1,000 Genomes Project reference panels, together with advances in statistical methodology, have allowed for the imputation of genotypes at single nucleotide polymorphism (SNP) markers that are untyped in a cohort or case-control study. These imputation procedures facilitate the interpretation and meta-analyses of genome-wide association studies. A natural question when implementing these procedures concerns how best to take into account uncertainty in imputed genotypes. Here we compare the performance of the following three strategies: least-squares regression on the "best-guess" imputed genotype; regression on the expected genotype score or "dosage"; and mixture regression models that more fully incorporate posterior probabilities of genotypes at untyped SNPs. Using simulation, we considered a range of sample sizes, minor allele frequencies, and imputation accuracies to compare the performance of the different methods under various genetic models. The mixture models performed the best in the setting of a large genetic effect and low imputation accuracies. However, for most realistic settings, we find that regressing the phenotype on the estimated allelic or genotypic dosage provides an attractive compromise between accuracy and computational tractability. *Genet. Epidemiol.* 35:102–110, 2011. © 2011 Wiley-Liss, Inc.

Key words: GWAS; genotype imputation; mixture models

INTRODUCTION

The shared ancestry of chromosomes in a population results in haplotype stretches shared by different individuals. Making use of these shared haplotype stretches, and thereby accounting for the correlation of alleles at nearby markers (linkage disequilibrium, LD), statistical algorithms can make inferences about unobserved alleles. To estimate a missing allele at a specific single nucleotide polymorphism (SNP) on a haplotype, these algorithms compare flanking markers with those from other haplotypes in the sample to find appropriate "template" or reference haplotypes to inform an estimate of the missing allele.

Recently there has been considerable interest in the imputation of missing genotype data for the analysis of genome-wide association (GWA) studies. The availability of panels of extensively genotyped reference samples, such as those from The International HapMap Project [HapMap; International HapMap Consortium, 2007] and now the 1,000 Genomes Project, has allowed for the indirect measurement of SNP genotypes that were not directly typed in a genetic association study but typed in the reference samples. This strategy has aided the discovery of multiple loci associated with disease [e.g. Barrett et al., 2008; Scott et al., 2007; The Wellcome Trust Case Control Consortium, 2007] or quantitative traits [Lettre et al., 2008; Loos et al., 2008; Willer et al., 2008]. For example, in Willer et al. [2008], the *LDLR* (cholesterol receptor) signal was detected only after imputation was performed, since the associated variant (rs6511720) was poorly tagged in samples genotypes with the Afymetrix 500K array set (maximum $R^2 \approx 0.21$).

This imputation-based mapping protocol is a 2-step process. First, unmeasured genotypes are imputed in the GWA data. Then, imputed genotypes are tested for association with phenotypes. Multiple methods exist for imputing genotypes from population genetic data [Browning and Browning, 2007; Greenspan and Geiger, 2004; Li et al., 2006; Marchini et al., 2007; Scheet and Stephens, 2006; Stephens and Scheet, 2005; for a review see Browning [2008]. Here we focus on the second step, testing the imputed genotypes for association with a trait of interest.

Specifically, we aim to evaluate the relative performance of several strategies for analyzing the distribution of imputed genotypes in downstream analyses. One summary of these probabilities comes from imputing a "best-guess" genotype for each individual, which corresponds to the marginal mode of the posterior distribution of the unmeasured genotype. This approach ignores the uncertainty in the imputed

genotype. When imputation is accurate, the correspondence between the true and imputed genotype is strong and an analysis of the imputed genotypes might result in little loss in power compared with the true genotypes. However, if imputation accuracy is low there may be a weak correlation between the true genotypes and the guesses, which will mask any real association between genotype and phenotype.

We also consider two approaches that attempt to account for this uncertainty. The first of these uses the mean of the distribution of imputed genotypes, which corresponds to an expected allelic or genotypic count, or "dosage", for each individual. This approach may do well, relative to using the "best guess" genotype, when there is some uncertainty about the true genotype, since it retains more of the available information, differentiating genotypes that were imputed with high confidence from those with greater uncertainty.

A final approach uses mixture regression models to take full advantage of the individual genotype posterior probabilities. This approach should be superior when there is uncertainty in the imputed genotypes, and information about the relationship between genotype and phenotype is not well summarized by a single average genotype. For example, this may occur when the posterior probabilities are high for the two homozygote genotypes, yet an average or allelic dosage would indicate that unmeasured genotype was a heterozygote.

We find that for most realistic settings of GWA studies, such as modest genetic effects, large sample sizes and high average imputation accuracies, the strategy of regressing the phenotype on the genetic dosages provides adequate performance. In fact, for these settings, small gains from using the full mixture models are rendered negligible by the increased model complexity and associated "cost" of estimating additional parameters. However, when the effect size is large, and imputation accuracy sufficiently poor, we demonstrate an increase in power when utilizing all available information in the posterior distribution in the form of mixture models.

METHODS

OVERVIEW

To simulate data from realistic cohort-based association studies, we first generated dense genotype data from a coalescent model. Then, conditional on these genotypes, we simulated quantitative trait data for all individuals in each cohort. In order to mimic the marker density from a GWA study, we masked a fraction of the SNPs and then imputed these genotypes, conditional on a set of simulated reference haplotypes and the remaining observed SNPs. Finally, we performed analyses to test for association between imputed genotype and phenotype.

SIMULATIONS

Genotype data. For each of 100 one-megabase (1-Mb) regions, we simulated 10,000 chromosomes from a coalescent model that mimics LD in real data, accounts for variation in local recombination rates, and models population history consistent with the HapMap CEU and YRI analysis panels [Schaffner et al., 2005].

For each 1-Mb region, we then took a random subset of 120 simulated chromosomes to generate a region-specific "pseudo HapMap". We randomly paired (assuming Hardy-Weinberg equilibrium) a random subset of 2,000

chromosomes of the remaining 9,880 chromosomes to create 1,000 diploid pseudo individuals.

For the simulated HapMap data, polymorphic sites were ascertained and thinned to match the corresponding (CEU) Phase II HapMap International HapMap Consortium [2007] marker density, allele frequency spectrum and LD patterns, resulting in $\approx 1,000$ SNPs for each region for the panel of 120 HapMap chromosomes. Based on the thinned HapMap panel, we selected a set of 100 tagSNPs for each region that included the 90 tagSNPs with the largest number of proxies and 10 additional SNPs picked at random among those remaining [Carlson et al., 2004]. The tagSNP selection approach taken above resulted in tagSNP sets that captured $\approx 78\%$ of the common variants (MAF $> 5\%$) in the simulated CEU HapMap, similar to the observed performance of the Illumina HumanHap300 Beadchip SNP genotyping platform. The genotypes at these 100 tagSNPs constituted the observed data for each simulated sample.

QUANTITATIVE TRAIT

We generated phenotype values on each of the n individuals for a large and small sample ($n = 1,000, 50$), conditional on their simulated genotypes. We simulated trait values separately for four genetic models, with varying degrees of dominance, and also for a null model where genotypes and phenotypes were independent.

At each SNP, the genotype label (0, 1, 2) is represented by the count of an arbitrarily chosen allele. Table I contains a summary of notation for the frequencies and genetic effect sizes ("phenotypic deviations") of each genotype. Since allele frequency affects the power to detect phenotype association, we adjust the phenotypic deviations separately for each SNP, so that we may tabulate results over all SNPs. To accomplish this, we maintain constant genetic variance attributable to the marker V_G, which we calculate with the following formula from Equation [8.8] (p. 129) of Falconer [1989]:

$$V_G = 2pq[a + d(q - p)]^2 + [2pqd]^2, \tag{1}$$

where p and $q = 1-p$ are allele frequencies, and a and d are additive and dominance effects (see Table I). We report genetic variance as a percent of total phenotypic variation (heritability), fixing this at 2.8% for $n = 1,000$, and 59.8% for $n = 50$. These values were calculated so as to achieve approximately 90% power at type-I error of 5×10^{-5} when analyzing the simulated genotypes under an additive genetic model with equal allele frequencies of one-half.

We performed the above trait simulations for 83,327 SNPs in turn for the following genetic models: additive ($d = 0$); partially dominant ($d = (1/2)a$); dominant ($d = a$);

TABLE I. Genotype and phenotype values

	Genotype		
	A/A	A/a	a/a
Labels	0	1	2
Frequencies	q^2	$2pq$	p^2
Phenotypic deviation	$-a$	D	a

Genotype labels are the counts of an arbitrarily chosen allele. The phenotypic deviations are the deviations from the mean μ^* used in the simulation, and vary by SNP. [This table is adapted from Table 7.3 of Falconer, 1989, p 121.]

and over-dominant ($d = (6/5)a$), corresponding to a value for the heterozygote that is 10% greater than the difference between the two homozygotes.

To simulate trait data y_i^* for individual i $(1,\dots,n)$ at a single SNP, we used the following model:

$$y_i^* = \mu^* + (-a)I_{\{g_i=0\}} + (d)I_{\{g_i=1\}} + (a)I_{\{g_i=2\}} + \varepsilon_i, \qquad (2)$$

where g_i is the true genotype for individual i, a and d are chosen according to (1), the indicator variable $I_{\{A\}}$ is one if A is true and zero otherwise, and $\varepsilon_i \sim \mathscr{N}(0,1)$.

REAL GENOTYPE DATA

We also obtained data from a GWA study of Type II diabetes in individuals of European descent [FUSION; Scott et al., 2007]. In 538 control samples, additional genotyping was conducted in a region of chromosome 14. This resulted in 521 markers for which we had both imputed genotypes from the HapMap, as well as genotypes typed on a custom microarray from Illumina. Imputation in these data set was carried out with MaCH 1.0 [Li et al., 2010], using the data from an Illumina 317 K microarray platform as "tag SNPs". Conditional on the typed genotypes at these 521 markers, we simulated quantitative trait data as above, keeping the genetic variance fixed to yield informative and interpretable summaries of power across multiple markers with varying allele frequencies. In addition to simulating phenotypes on the full set (538 individuals), we also simulated a larger effect on a subset of 50 individuals, selected at random. We repeated these simulations 100 times to obtain simulated phenotypes for 52,100 SNPs (100×521), for both "large" and small sample sizes.

GENOTYPE IMPUTATION

To obtain posterior probabilities and imputed genotypes, (Fig. 1) we used the software package fastPHASE [Scheet and Stephens, 2006]. For each simulated region, we fit the LD model to the reference chromosomes only, and then applied this fitted model to the pseudo individuals in the simulated cohort. (For convenience we set the number of haplotype clusters K to be 20.) We assess imputation accuracy with the square of the Pearson correlation coefficient between the true and best-guess genotypes (R^2), which is more informative about power at different allele frequencies than a simple genotype imputation error rate measure. For our simulations, the median $\hat{R}^2$ for these data was 0.90 and the mean was 0.75.

REGRESSION ANALYSIS

We used regression analysis to test the effectiveness of multiple summaries of the imputed genotypes. Let p_{ki} denote the conditional ("posterior") probabilities for the imputed genotypes of individual i $(1,\dots,n)$, where k $(0,1,2)$ indexes the genotype by its label. We evaluated the performance of the following three summaries of the genotype probabilities conditional on the observed data:

1. Best guess—maximum a posteriori ("MAP") genotype;
2. Dosage—estimated (expected) allelic or genotypic counts; and
3. Posterior probabilities—probabilities of the three possible genotypes obtained from imputation.

For comparison, we also analyzed the true (simulated or typed) genotypes.

First we give the models used for ordinary least squares (OLS) regression. Then, we explain the use of mixture

models for regression. For each method, we consider both additive (1-parameter of 1-degree-of-freedom "1 df") and non-additive (2-parameter, 2 df) regression models for analysis. In what follows, let y_i denote the quantitative trait value for individual i at a SNP.

ORDINARY REGRESSION ON GENOTYPES AND ALLELIC DOSAGE

Additive. Let x_i represent a particular feature of the imputation procedure or the true genotype (g_i) at a SNP under consideration, i.e.

$$x_i = \begin{cases} \arg\max_{k\in\{0,1,2\}}\{p_{ki}\} & \text{best-guess genotype} \\ p_{1i} + 2p_{2i} & \text{allelic dosage} \\ g_i & \text{true genotype.} \end{cases}$$

The additive model is written as

$$y_i = \mu + \beta x_i + \varepsilon_i, \qquad (3)$$

where $\varepsilon \sim \mathscr{N}(0,\sigma^2)$, independently for all i. We use OLS regression to test the null hypothesis $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$. To evaluate significance, we compute an F-statistic.

Non-additive. Under a non-additive model, we expand x_i to be composed of two components $(x_i^{(1)}, x_i^{(2)})$ as follows:

$$(x_i^{(1)}, x_i^{(2)}) = \begin{cases} (I_{\{x_i=1\}}, I_{\{x_i=2\}}) & \text{best-guess genotype} \\ (p_{1i}, p_{2i}) & \text{genotypic dosage} \\ (I_{\{g_i=1\}}, I_{\{g_i=2\}}) & \text{true genotype.} \end{cases}$$

We write the dominance model as

$$y_i = \mu + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \varepsilon_i, \qquad (4)$$

where $\varepsilon_i \sim \mathscr{N}(0,\sigma^2)$, as above. Again we evaluate the null hypothesis that there is no effect for any genotype, i.e. $H_0 := \beta_1 = 0, \beta_2 = 0$ vs. $H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$. We apply OLS regression and compute an F-statistic.

MIXTURE OF REGRESSION MODELS

To investigate the approach of multiple-imputation, we fit a mixture of regression models to the phenotype data and posterior genotype probabilities. The composite regression model may be written as

$$y_i = \sum_{k=0}^{2} p_{ki} f_k(\mu, \beta, \varepsilon_i), \qquad (5)$$

where the regression function $f_k(\cdot)$ is a function of the assumed genetic model, i.e. additive or non-additive (see below).

For each assumed model below, we construct likelihood ratio statistics to test for statistical significance. To estimate the parameters (μ, β), we maximize the log-likelihood function using the Nelder-Mead Simplex Method [Nelder and Mead, 1965], implemented in the R package *optim*.

Additive. Under an assumption of additivity of the allelic effects, the regression function $f_k(\cdot)$ is

$$f_k(\mu, \beta, \varepsilon_i) = \begin{cases} \mu + \varepsilon_i, & k = 0 \\ \mu + \beta + \varepsilon_i, & k = 1 \\ \mu + 2\beta + \varepsilon_i, & k = 2, \end{cases} \qquad (6)$$

where $\varepsilon_i \sim \mathscr{N}(0,\sigma^2)$. To test the hypothesis $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$, we construct a likelihood ratio test.

Non-additive. Relaxing the assumption of additivity (allowing for dominance) of the allelic effects, we expand β to be (β_1, β_2), and the regression function $f_k(\cdot)$ is

$$f_k(\mu, \beta_1, \beta_2, \varepsilon_i) = \begin{cases} \mu + \varepsilon_i, & k = 0 \\ \mu + \beta_1 + \varepsilon_i, & k = 1 \\ \mu + \beta_2 + \varepsilon_i, & k = 2, \end{cases} \quad (7)$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. To test the hypothesis $H_0 : \beta_1 = 0$ $\beta_2 = 0$ vs. $H_1 : \beta_1 \neq 0$, or $\beta_2 \neq 0$, we construct a likelihood ratio test.

RESULTS

Here we present results on simulated phenotypes from both simulated and real genotype data, as well as imputed genotypes from standard software packages. For various settings (sample sizes, effect sizes, real and simulated genotypes, genetic models), we tabulate power results overall, as well as plot them by imputation accuracy and allele frequency (calculated from the full cohort of 1,000 individuals).

LARGE SAMPLE SIZE WITH SMALL EFFECTS

We computed power empirically, based on the analysis of ≈ 1 million null data sets (where there was no association between phenotype and genotype) from which we obtained empirical significance thresholds. Results from analyses of our various imputation strategies and regression models, for the large sample of 1,000 individuals in the simulated studies, are reported in Table II.

In general, there was a consistent gain in performance achieved from using the dosage summaries or mixture models in comparison to using the best guess genotypes. This improvement was larger for the two-parameter regression models, regardless of the underlying genetic model, with absolute gains in power of $\approx 14\%$. For additive or one-parameter models, the average gain was more modest (3–4%). All differences between the dosage and mixture model strategies were small (<2%).

We also examined the effect of imputation accuracy and allele frequencies on the power to detect association in Figure 2. We summarized accuracy at each SNP with the square of the Pearson correlation coefficient between the imputed and true genotypes (coded as 0, 1, or 2), which we refer to as R^2.

When the accuracy is high ($R^2 > 0.9$), using the best-guess genotype from the imputation procedure results in little loss of power. The gain from using a dosage or mixture model is greatest at intermediate accuracies, since posterior probabilities are informative about the underlying genetic variation, even if they do not allow accurate "best-guess imputation" of genotypes. For all three strategies, at low imputation accuracies, the lines of the additive regression models converge, so do the lines of the dominant regression models.
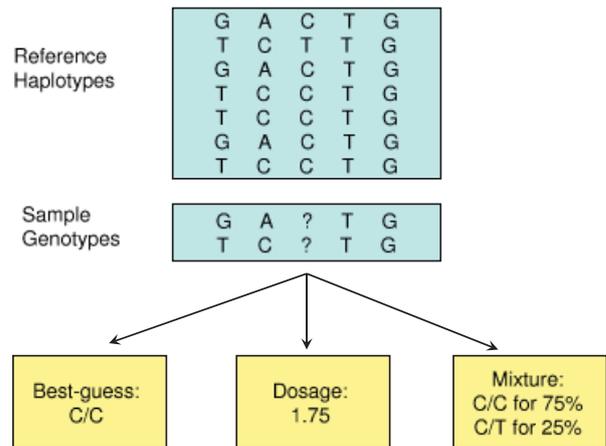

Fig. 1. Example of posterior probability summaries. Here we present a didactic illustration of the three summaries of the full posterior probabilities for imputed genotypes. From the set of Reference Haplotypes, the missing genotype (denoted with two ? symbols) in the sample genotypes can be inferred. Based on the reference, the first sample haplotype would consist of a C at the missing position, since all three similar haplotypes in the reference set have a C here. For the second sample haplotype, three-fourths of the similar haplotypes in the reference set consist of a C; and one consists of a T at that position. Therefore, the "expected" dosage would be 1.75. And the only "possible" genotypes, based completely on the reference, would be C/C and C/T, expected probabilities given.

TABLE II. Power results for large sample size and small effects

Analysis strategy Imputation summary/regression model	Simulated trait			
	Additive $(d = 0)$	Partially dominant $(d = (1/2)a)$	Dominant $(d = a)$	Over-dominant $(d = (6/5)a)$
Best-guess/1 df	0.635	0.599	0.478	0.435
Best-guess/2 df	0.466	0.463	0.448	0.449
Dosage/1 df	0.660	0.620	0.489	0.447
Dosage/2 df	0.603	0.598	**0.588**	**0.588**
Mixture/1 df	**0.668**	**0.628**	0.499	0.456
Mixture/2 df	0.604	0.600	0.587	**0.588**
True/1 df	0.897	0.865	0.730	0.683
True/2 df	0.708	0.711	0.706	0.709

The "Analysis Strategy" specifies the combination of imputation summary (e.g. best guess, dosage, or mixture model) and whether the regression model assumes a strict additive model (Equation (3); "1 df") or allows for dominance (Equation (4); "2 df"). Results are based on a cohort of 1,000 individuals. Power was computed at a fixed type-I error rate (α) of 5×10^{-5}, based on empirical quantiles from analysis of 916,597 "null" data sets, with a trait simulated independent of genotype. Quantitative traits were simulated to have constant genetic variance of 2.8% heritability.
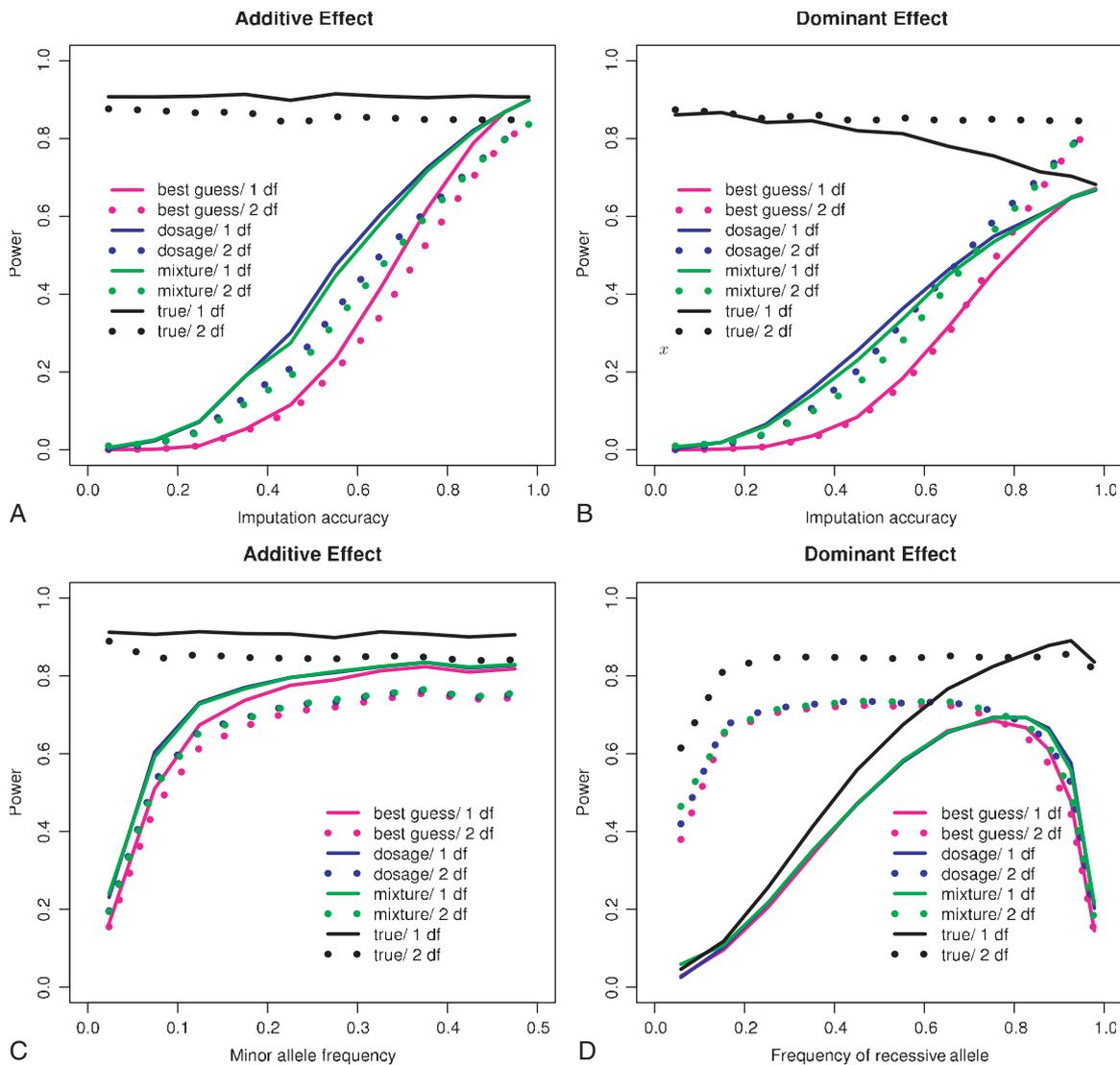
Fig. 2. Power vs. accuracy and allele frequency for large sample size and small effects. For each summary and the true genotypes, both an additive (solid line) and dominant (dotted line) model were analyzed. (A) and (C) are based on data simulated with an additive effect; (B) and (D) are based on data simulated under a model of complete dominance. Power was computed at a fixed type-I error rate (α) of 5×10^{-5}. The sample size was 1,000. TOP: Power is plotted against R^2, a measure of imputation accuracy. BOTTOM: Power is plotted against allele frequency. (A) Power vs. R^2 with an additive effect; (B) power vs. R^2 under complete dominance; (C) power vs. frequency of minor allele with an additive effect; and (D) power vs. frequency of dominant allele under complete dominance.

An important factor in overall power summaries, such as those in Tables II and III (below), is the allele frequency distribution of SNPs present in the reference panel, at which genotypes are being imputed in the study samples, since the tables are constructed with averages over all SNPs. In Figures 2C and 3C, where phenotypes were simulated from an additive genetic model, powers for all regression models increase substantially when minor allele frequencies are relatively low. This may reflect the relative difficulty of accurate imputation at SNPs with a lower MAF. (Under the correct additive model, power for the true genotypes is unaffected, since we attempted to make power independent of allele frequency for the purposes of aggregating results across SNPs for general comparisons among analysis strategies; see Methods.) For data simulated

under a dominant genetic model, methods that assume the correct dominant model for analysis are superior at a greater range of allele frequencies.

SMALL SAMPLE SIZE WITH LARGE EFFECTS

For SNPs with modest genetic effects, as above, there is little gain from the increased computational demands of applying mixture models for the analyses. To examine a scenario where the mixture models might offer an advantage, we repeated the above simulations with larger genetic effects (and thus smaller sample sizes so that power was below 100%). This situation might be found in expression quantitative trait loci (eQTL) mapping studies, for example. These results are in Table III.

TABLE III. Power results for large effects and small sample size

Analysis strategy Imputation summary/regression model	Simulated trait			
	Additive $(d = 0)$	Partially dominant $(d = (1/2)a)$	Dominant $(d = a)$	Over-dominant $(d = (6/5)a)$
Best-guess/1 df	0.701	0.688	0.582	0.546
Best-guess/2 df	0.682	0.670	0.629	0.636
Dosage/1 df	0.755	0.743	0.629	0.590
Dosage/2 df	0.745	0.736	0.702	0.707
Mixture/1 df	**0.850**	**0.837**	0.721	0.686
Mixture/2 df	0.829	0.828	**0.805**	**0.810**
Truth/1 df	0.916	0.911	0.802	0.767
Truth/2 df	0.913	0.910	0.873	0.890

The "Analysis Strategy" specifies the combination of imputation quantity/summary (e.g. best guess, dosage, or mixture model) and whether the regression model allows for deviations from a strict additive model. Results are based on a cohort of 50 individuals. Power was computed at a fixed type-I error rate (α) of 5×10^{-5}, based on based on empirical quantiles from analysis of 916,597 "null" data sets, with a trait simulated independent of genotype. Quantitative traits were simulated to have constant genetic variance (of 59.8% heritability), given the genetic model and allele frequencies at each SNP.

Here, the advantage of applying mixture models is apparent, with average power gains of 10–12%. The contrast is greater at lower imputation accuracies (top row of Fig. 3) and is maintained even when we applied the incorrect additive regression model to data simulated with a strong dominant effect (Fig. 3B) (i.e. the green solid line is well above the blue dotted line at modest accuracies).

It is worth noting that although we attempted to simulate phenotypes so that results may be tabulated across allele frequencies, by keeping heritability constant per Equation (1), this does not guarantee that power will be independent of allele frequency. Heritability represents, in some sense, the amount of information about the phenotype and genotype relationship. How that relates to power is not completely predictable and will depend on additional factors, such as analysis methods, genetic model (e.g. dominance here), and particularly imputation accuracy. However, even for the true genotypes, it is difficult to calibrate power at low allele frequencies in samples of finite size. For example, in Figures 2D and 3D, power is reduced at low frequencies of the dominant allele. This is due to the requirement of having a homozygote—a rare genotype at low allele frequency and thus less likely to be observed near its expected proportion—for a shift in phenotypic mean; e.g., this phenomenon is more pronounced in Figure 3D where the sample size is 50. In fact, it is for this reason that we included results for true genotypes (where the "correction" for allele frequency, etc., was not perfect).

COMPUTATION

The mixture-model-based procedures were considerably more computationally demanding, based on our implementations in the R software package. Per-marker run times for the mixture-models averaged approximately 4 sec for 1-df (about 300 times longer than for the best-guess and dosage methods) and 20 sec for 2-df regression models. However, calculations for methods applied in this study can be conducted in parallel. We estimate that an application of mixture models to poorly imputed SNPs in a GWA study could be completed in a couple of days using tens of CPUs.

REAL DATA WITH LARGE AND SMALL EFFECTS

We confirmed the general applicability of our results to real genotype data, by applying our methods to 538 control samples from a GWA case-control study of Type II diabetes (FUSION). We studied the following two scenarios: (1) all 538 samples and a modest effect (single-marker heritability of 4.3%); and (2) small sample size of 50 individuals and a large effect (single-marker heritability of 59.8%). To examine the phenomenon of seeing greatly increased power for the mixture models at sites with poor imputation accuracy, we report results for small sample size by low imputation accuracy ($R^2 < 0.56$) and "high" accuracy ($R^2 \geq 0.56$). (Due to the constraints of the real data, there does not exist a full spectrum of allele frequencies for plots by allele frequency. The cutoff of 0.56 was chosen based on a visual examination of Fig. 3.) In all scenarios, the power from using mixture models equals or exceeds those for the dosage and best-guess summaries, although only the scenario of low imputation accuracy and large effects show a pronounced difference. Results are displayed in Table IV.

DISCUSSION

Several software packages have been developed to impute and test SNPs that were not typed directly, such as BIMBAM [Servin and Stephens, 2007], IMPUTE [Marchini et al., 2007], MaCH [Li et al., 2009, 2010], and Beagle [Browning and Browning, 2009]. Two of these methods (BIMBAM and IMPUTE) assess association between genotype and phenotype with a Bayes Factor. We do not consider the Bayesian approach here, but this is discussed by Guan and Stephens [2008].

Multiple factors will impact power of imputation-based strategies for the analysis of GWA studies, including differences in the patterns of LD and allele frequencies between the study and reference populations. However, for the single-marker analyses examined in our study, the impact of these factors can be measured via their effect on imputation accuracy, since the missing (unmeasured) genotypes are the quantities of interest for analysis. Different imputation algorithms will lead to slightly differential accuracies. However, our aim here was not to
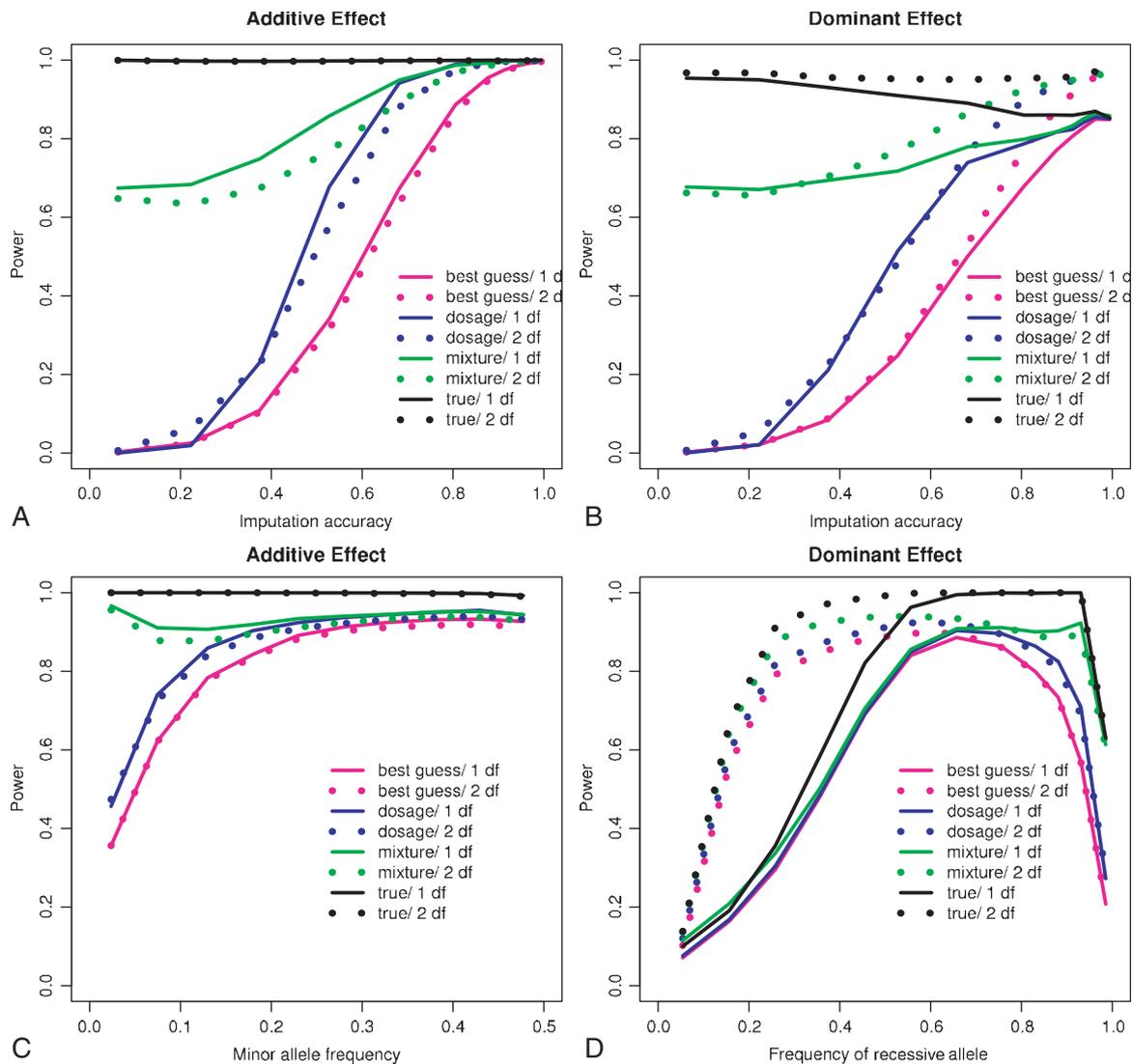
Fig. 3. Power vs. accuracy and allele frequency for small sample size and large effects. Power was computed at a fixed type-I error rate (α) of 5×10^{-5}. The sample size was 50. For each summary and the true genotypes, both an additive (solid line) and dominant (dotted line) model were analyzed. (A) and (C) are based on data simulated with an additive effect; (B) and (D) are based on data simulated under a model of complete dominance. TOP: Power is plotted against R^2, a measure of imputation accuracy. BOTTOM: Power is plotted against allele frequency: (A) power vs. R^2 with an additive effect; (B) power vs. R^2 under complete dominance; (C) power vs. frequency of minor allele with an additive effect; and (D) power vs. frequency of dominant allele under complete dominance.

compare these accuracies but to condition on the sorts of accuracies that might be expected from typical marker densities and patterns of LD.

The SNP density targeted in our simulations was motivated by analyses of existing GWA studies. Increased densities should result in more information about LD to increase imputation accuracies. For this reason, we plotted results by imputation accuracies in addition to the tables which integrate over the distributions of LD patterns and allele frequencies. For the same reason, our results should be applicable to imputation from low-coverage sequencing data. Although the distributions of allele frequencies of interrogated SNPs will shift to lower values, and imputation accuracies may vary in a manner different from those encountered in this study, our results plotted by these features (frequency and accuracy) should apply to other

raw data sources. (Our tabulated summaries may in fact change under these different conditions, since the results are integrated over particular distributions of allele frequencies and accuracies, dependent on the simulations and imputation methods employed herein.)

We applied methods to a sample size of 1,000 individuals. While this size is somewhat smaller than for some GWA studies, and much smaller than associated meta-analyses, it is sufficiently large to illustrate comparisons of methods for effect sizes that correspond to intermediate power to detect association. Larger sizes, with similarly sized effects, will simply result in increased power regardless of methods. Smaller sample sizes will require stronger genetic effects for there to exist sufficient power to detect association. Examples of such scenarios may come from studies of pharmacogenetics or mapping eQTLs.

TABLE IV. Power results for various effect and sample sizes: application to real data

| Scenario Imputation summary/regression model | $n = 538$, modest effect | | $n = 50$, large effect | | | |
| | | | $R^2 < 0.56$ | | $R^2 \geq 0.56$ | |
	Additive $(d = 0)$	Dominant $(d = a)$	Additive $(d = 0)$	Dominant $(d = a)$	Additive $(d = 0)$	Dominant $(d = a)$
Best-guess/1 df	0.708	0.244	0.188	0.043	0.983	0.447
Best-guess/2 df	0.616	0.573	0.148	0.112	0.975	0.710
Dosage/1 df	0.716	0.249	0.297	0.057	0.983	0.448
Dosage/2 df	0.619	0.587	0.305	0.214	0.976	0.723
Mixture/1 df	**0.720**	0.255	**0.558**	0.220	**0.992**	0.474
Mixture/2 df	0.621	**0.592**	0.469	**0.396**	0.983	**0.752**
Truth/1 df	0.798	0.287	1.000	0.406	1.000	0.478
Truth/2 df	0.710	0.674	0.999	0.544	0.999	0.766

Here we give power computed at a fixed type-I error rate (α) of 5×10^{-5}, calculated from theoretical distributions. We confirmed that type-I error was calibrated at the 10^{-3} level (results not shown). For each summary of the imputed genotypes (best-guess, dosage, and mixture) plus the true genotypes, we apply 2 regression models, assuming additivity (Equation (3); "1 df") and dominance (Equation (4); "2 df"). For the small sample of 50 individuals, we report power for SNPs imputed at low imputation accuracy ($R^2 < 0.56$) and high accuracy ($R^2 \geq 0.56$). For the full sample ($n = 538$), genetic variance was fixed at 4.3% and for the small sample ($n = 50$), it was 59.8%.

Here, we have made no attempt to model the correlation of genotypes among SNPs during analysis. To detect interactions among genotypes at nearby SNPs, it may be beneficial to model this dependence during imputation and analysis. The imputation procedures mentioned above may obtain correlated genotypes by sampling entire chromosomes of untyped SNPs, instead of the data at each SNP, marginally.

It may be possible to do better in such a setting by using genuine "multiple imputation" methods. However, in our setting, by applying a mixture of regression models, we hope to capture a range of possible phenotype-genotype relationships, and the gain from multiple imputation over the mixture model should not be large. Therefore, we felt that the mixture model provided a close approximation to an optimal analysis procedure.

In our most relevant comparisons with modest effects and large sample sizes, use of the dosage summaries was as powerful as using the mixture model methods, at a fraction of the computational cost. The exception to this result is apparent only at SNPs with very large genetic effects. In such situations of large effects, most methods will be effective at detecting an association. This difference is most pronounced at poorly imputed SNPs. In practice, many researchers routinely exclude results from poorly imputed SNPs, such as those below an R^2 threshold of, say, 30%. Application of this quality-control filter to our results would tend to mitigate (tabulated) differences in power between the mixture and standard regression methods in the setting of large effect sizes. In fact, it may be fruitful, in some cases, to devote additional computational resources to some of these SNPs, such as application of mixture models. However, for the majority of settings and effect sizes detected and verified in GWA studies, use of dosage quantities appears to be effective and efficient to account for the uncertainty in the imputed genotypes.

ACKNOWLEDGMENTS

REFERENCES

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Belgian-French IBD Consortium, Wellcome Trust Case Control Consortium, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. Nat Genet 40:955–962.

Browning SR. 2008. Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet 124: 439–450.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084–1097.

Browning BL, Browning SR. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84:210–223.

Carlson C, Eberle M, Rieder M, Yi Q, Kruglyak L, Nickerson D. 2004. Selecting a maximally informative set of single-nucleotide poly-morphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106–120.

Falconer DS. 1989. Introduction to Quantitative Genetics (3rd edition). New York: Longman Scientific & Technical.

Greenspan G, Geiger D. 2004. Model-based inference of haplotype block variation. J Comput Biol 11:493–504.

Guan Y, Stephens M. 2008. Practical issues in imputation-based association mapping. PLoS Genet 4:e1000279.

International HapMap Consortium. 2007. A second generation human haplotypemap of over 3.1 million SNPs. Nature 449:851–861.

Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, Eyheramendy S, Voight BF, Butler JL, Guiducci C, Illig T, Hackett R, Heid IM, Jacobs KB, Lyssenko V, Uda M, Boehnke M, Chanock SJ, Groop LC, Hu FB, Isomaa B, Kraft P, Peltonen L, Salomaa V, Schlessinger D, Hunter DJ, Hayes RB, Abecasis GR, Wichmann HE, Mohlke KL, Hirschhorn JN. 2008. Identification of ten loci associated with height highlights new biological pathways in human growth. Nat Genet 40:584–591.

Li Y, Ding J, Abecasis GR. 2006. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. Am J Hum Genet 79:S2290.

Li Y, Willer C, Sanna S, Abecasis GR. 2009. Genotype imputation. Annu Rev Genom Hum Genet 10:387–406.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34:816–834.

Loos RJF, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, Inouye M, Freathy RM, Attwood AP, Beckmann JS, Berndt SI, Bergmann S, Bennett AJ, Bingham SA, Bochud M, Brown M, Cauchi S, Connell JM, Cooper C, Smith GD, Day I, Dina C, De S, Dermitzakis ET, Doney ASF, Elliott KS, Elliott P, Evans DM, Farooqi IS, Froguel P, Ghori J, Groves CJ, Gwilliam R, Hadley D, Hall AS, Hattersley AT, Hebebrand J, Heid IM, KORA, Herrera B, Hinney A, Hunt SE, Jarvelin MR, Johnson T, Jolley JDM, Karpe F, Keniry A, Khaw KT, Luben RN, Mangino M, Marchini J, McArdle WL, McGinnis R, Meyre D, Munroe PB, Morris AD, Ness AR, Neville MJ, Nica AC, Ong KK, O'Rahilly S, Owen KR, Palmer CNA, Papadakis K, Potter S, Pouta A, Qi L, Nurses' Health Study, Randall JC, Rayner NW, Ring SM, Sandhu, Scherag A, Sims MA, Song K, Soranzo N, Speliotes EK, Diabetes Genetics Initiative, Syddall HE, Teichmann SA, Timpson NJ, Tobias JH, Uda M, The SardiNIA Study, Vogel CIG, Wallace C, Waterworth DM, Weedon MN, The Wellcome Trust Case Control Consortium, Willer CJ, FUSION, Wraight VL, Yuan X, Zeggini E, Hirschhorn JN, Strachan DP, Ouwehand WH, Caulfield MJ, Samani NJ, Frayling TM, Vollenweider P, Waeber G, Mooser V, Deloukas P, McCarthy MI, Wareham NJ, Barroso I. 2008. Common variants near mc4r are associated with fat mass, weight and risk of obesity. Nat Genet 40:768–775.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913.

Nelder JA, Mead R. 1965. A simplex algorithm for function minimization. Comput J 7:308–313.

Schaffner S, Foo C, Gabriel S, Reich D, Daly M, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15:1576–1583.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629–644.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanana TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316:1341–1345.

Servin B, Stephens M. 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3:e114.

Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449–462.

The Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, Strait J, Duren WL, Maschio A, Busonero F, Mulas A, Albai G, Swift AJ, Morken MA, Narisu N, Bennett D, Parish S, Shen H, Galan P, Meneton P, Hercberg S, Zelenika D, Chen WM, Li Y, Scott LJ, Scheet PA, Sundvall J, Watanabe RM, Nagaraja R, Ebrahim S, Lawlor DA, Ben-Shlomo Y, Davey-Smith G, Shuldiner AR, Collins R, Bergman RN, Uda M, Tuomilehto J, Cao A, Collins FS, Lakatta E, Lathrop GM, Boehnke M, Schlessinger D, Mohlke KL, Abecasis GR. 2008. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40:161–169.

APPENDIX A: EFFECT SIZES FOR SIMULATIONS

Phenotypes were simulated as described in Methods, using Equations (1) and (2). Here, we focus on the values used for the more realistic scenario of a larger samples size (1,000), i.e. a constant genetic variance with heritability of 2.8%, chosen for adequate power to facilitate comparisons among methods. In Figure A1 we show the actual effect sizes—the values for a and d used in Equation (2)—as they vary with allele frequency. Note the frequency of the *recessive* allele is plotted on the horizontal axis, and thus for the purely additive model (no dominance) the effect size is symmetric about an allele frequency of 0.5.
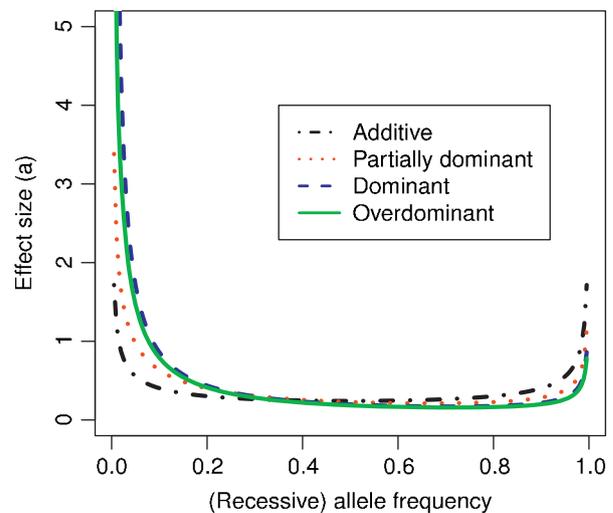

Fig. A1. Summary of effect sizes for phenotype simulations. Values for the effect size (a) are plotted against allele frequencies of the recessive allele (allele "A" in Table I). Values of d are given as in Tables II and III, i.e. 0 (Additive), (1/2)a (Partially dominant), a (Dominant), and (6/5)a (Overdominant).