The Relationship between Imputation Error and Statistical Power in Genetic Association Studies in Diverse Populations

Lucy Huang,[1] Chaolong Wang,[1] and Noah A. Rosenberg[1,2,3,*]

Genotype-imputation methods provide an essential technique for high-resolution genome-wide association (GWA) studies with millions of single-nucleotide polymorphisms. For optimal design and interpretation of imputation-based GWA studies, it is important to understand the connection between imputation error and power to detect associations at imputed markers. Here, using a 2 × 3 chi-square test, we describe a relationship between genotype-imputation error rates and the sample-size inflation required for achieving statistical power at an imputed marker equal to that obtained if genotypes at the marker were known with certainty. Surprisingly, typical imputation error rates (~2%–6%) lead to a large increase in the required sample size (~10%–60%), and in some African populations whose genotypes are particularly difficult to impute, the required sample-size increase is as high as ~30%–150%. In most populations, each 1% increase in imputation error leads to an increase of ~5%–13% in the sample size required for maintaining power. These results imply that in GWA sample-size calculations investigators will need to account for a potentially considerable loss of power from even low levels of imputation error and that development of additional genomic resources that decrease imputation error will translate into substantial reduction in the sample sizes needed for imputation-based detection of the variants that underlie complex human diseases.

The genotype-imputation strategy for case-control genetic association studies provides an economical way of assessing many more genetic markers for disease association than have actually been measured in any particular association study.[1–5] In this approach, case and control individuals are first genotyped for markers densely spread across the human genome. The measured genotypes are then combined with high-resolution genotypic data from genomic databases for imputation of the genotypic status of study individuals at markers investigated in the database but not in the study sample. This imputation relies on the principle that two haplotypes identical in genotype at nearby SNP markers are likely to share intervening chromosomal stretches identically by descent. Thus, if a haplotype in a densely genotyped database sample is identical to a haplotype in a more sparsely genotyped study sample for markers that overlap between the study and the database, then one can impute the study haplotype with high resolution by copying the haplotype from the database.

Methods relying on genotype imputation have proven effective for identifying high-risk disease-associated genetic variants, in part because they dramatically increase the number of markers that can be directly tested for association in comparison to earlier tag-SNP designs.[6–10] However, the imputation strategy utilizes in its association tests estimated genotypes that are not known with certainty, and errors in imputed genotypes might potentially compromise the power of an imputation-based association test. For example, at a biallelic marker, consider a disease-susceptibility allele of small effect that has a true frequency of 0.3 in cases and 0.2 in controls. If the probability that imputation recovers the true allele is 0.9, then the frequency of the disease allele among *imputed* genotypes will be $(0.3)(0.9) + (0.7)(0.1) = 0.34$ in cases and $(0.2)(0.9) + (0.8)(0.1) = 0.26$ in controls. Imputation error converts an allele-frequency difference of $0.3 - 0.2 = 0.1$ between cases and controls into a smaller difference of $0.34 - 0.26 = 0.08$. As a result, for the imputed genotypes, one might require a larger sample size in order to determine that allele frequencies differ between cases and controls, as compared to the smaller sample size that would be required if the true genotypes were known.

Although recent studies have found that imputation error rates are generally low,[11–15] it is possible that even low error rates could have considerable effects on downstream analyses. How does the error inherent in genotype imputation reduce the power of an association study when alleles at the true disease SNP are imputed rather than known? An answer to this question is important to the design and interpretation of imputation-based association studies. Relating imputation error and power would assist investigators in calculating sample sizes required for detecting disease variants at loci whose genotypes are imputed, as well as in determining whether imputation studies in particular populations are likely to be underpowered. Additionally, a relationship between imputation error and power would aid in the development of resources for genomic studies. For example, use of such a relationship could assist in the identification of populations in whom existing resources produce high error rates that limit the potential for practical mapping of risk variants with imputation strategies.

The problem of connecting imputation error to power is similar to a corresponding problem in the context of tag SNPs. In the imputation context, the loss of information

[1]Center for Computational Medicine and Bioinformatics; [2]Department of Human Genetics; [3]Life Sciences Institute
University of Michigan, 2017 Palmer Commons, 100 Washtenaw Ave., Ann Arbor, MI 48109, USA
*Correspondence: rnoah@umich.edu

DOI 10.1016/j.ajhg.2009.09.017.

as a result of imputation error at a disease-susceptibility locus can obscure the association between the locus and disease. In the tag-SNP context, the loss of information as a result of using a tag SNP rather than the true disease SNP has an analogous effect. In both situations, missing information about the correct genotypes at the true disease-susceptibility locus contributes to a loss of power for detecting disease association.

For the tag-SNP context, consider two loci, a SNP causally associated with disease and a nearby tag SNP. If the r^2 correlation statistic for linkage disequilibrium (LD) between the tag SNP and the disease SNP is equal to c, then a chi-square test statistic for disease association at the true disease SNP in a case-control sample of total size N has approximately the same asymptotic distribution under the alternative hypothesis of disease association as the corresponding chi-square statistic at the tag SNP in a case-control sample of size N/c.[16] Thus, the "sample-size inflation factor" required in using the tag SNP in an association study rather than the true disease SNP is ~$1/c$.

Motivated by this result, investigators have proposed multiple versions of an r^2 correlation statistic between the imputed genotypes at a SNP and the true genotypes.[17–19] Such statistics, which are sometimes used for identifying markers imputed with high accuracy in imputation-based genome-wide association (GWA) studies,[6,20] have been viewed as conceptually analogous to the r^2 statistic for LD between a tag-SNP and a disease SNP, but they have not been shown to be mathematically equivalent to it. In the imputation context for a biallelic SNP with alleles A and B, the correlation between true and imputed genotypes is a function of a 3 × 3 table, in which each of three possible true genotypes (AA, AB, BB) has one of three possible imputations. In the tag-SNP context, however, if the disease SNP has alleles A and B and the tag SNP has alleles C and D, then the corresponding table is a 2 × 2 table, containing entries for the counts of the four possible haplotypes (AC, AD, BC, BD). Although the close analogy between the tag SNP and imputation contexts suggests that the relationship between imputation error and power is similar to that observed between power and LD with a tag SNP, at present the connection between imputation r^2 statistics and power remains informal.

Here, in order to investigate the mathematical relationship between imputation error and power, we adapt a method developed for evaluating the relationship between *genotyping* error and power.[21,22] Our approach does not use an r^2 statistic, and unlike the inflation factor in the tag-SNP context, which depends only on the LD between the tag and disease SNPs, the corresponding inflation factor in the imputation context is a function of nine parameters. Consider two 2 × 3 chi-square tests of association, examining the relationship between the three possible genotypes of a biallelic marker and case-control status. The first test uses the true genotypes of the marker, whereas the second test uses genotypes measured with the possibility of imputation error. Suppose that k is the ratio of the number of controls to the number of cases. Denote by $\mathrm{MAF}_{controls}$ the frequency of the minor allele in controls, and by MAF_{cases} the frequency of this same allele in cases. Thus, $0 \leq \mathrm{MAF}_{controls} \leq 1/2$ and $0 \leq \mathrm{MAF}_{cases} \leq 1$. We label the minor allele in controls by A, the major allele in controls by B, genotype AA by 1, AB by 2, and BB by 3. For $i, j \in \{1, 2, 3\}$, we let ε_{ij} be the probability that genotype i is imputed as genotype j. Because $\sum_{j=1}^{3} \varepsilon_{ij} = 1$ for each i, only six error parameters must be considered: ε_{12}, ε_{13}, ε_{21}, ε_{23}, ε_{31}, and ε_{32}.

Gordon et al.[21] and Kang et al.[22] determined the relationship between the two 2 × 3 chi-square test statistics at a locus, showing that the test statistic for association between true genotype and disease in a sample of size N has the same asymptotic distribution as the test statistic for association between imputed genotype and disease in a sample of size Nf, in which $f \geq 1$ is a rational function of ε_{12}, ε_{13}, ε_{21}, ε_{23}, ε_{31}, ε_{32}, k, MAF_{cases}, and $\mathrm{MAF}_{controls}$ that represents the sample-size inflation factor. Thus, if a sample size of at least N is required for achieving a specified level of power when genotype is measured without error, then a sample size of at least Nf is required for achieving the same power when genotype is imputed with error. We use a special case of the formula for f, assuming $k = 1$, so that a study has an equal number of cases and controls. We also assume that Hardy-Weinberg proportions are satisfied separately in cases and controls. With these assumptions, the sample-size inflation factor due to imputation error can be written as $f = g/g^*$, defining g and g^* as in equations 1 and A.1 of Kang et al.[22] and matching our notation to that of Kang et al., with the substitutions $P_{01} = \mathrm{MAF}_{cases}^2$, $P_{02} = 2\mathrm{MAF}_{cases}(1 - \mathrm{MAF}_{cases})$, $P_{03} = (1 - \mathrm{MAF}_{cases})^2$, $P_{11} = \mathrm{MAF}_{controls}^2$, $P_{12} = 2\mathrm{MAF}_{controls}(1 - \mathrm{MAF}_{controls})$, and $P_{13} = (1 - \mathrm{MAF}_{controls})^2$.

To evaluate the sample-size inflation factor f at levels of imputation error appropriate for typical association studies, we first estimated the six error parameters by using genotypes of 426 individuals in 29 diverse populations. Employing reference panels of phased haplotypes based on ~2,000,000 SNPs in 210 HapMap Phase II individuals together with ~500,000 SNPs from a worldwide study,[23] we imputed individual genotypes at markers that were included in the reference data but not in the worldwide study. For each population, we repeated the imputations underlying Figure 7 of Huang et al.,[19] using the same procedure that was used by Huang et al.,[19] to obtain an imputed data set of 513 markers. This set consisted of probabilistic imputations relying on the subset of reference individuals that, among seven choices in the work of Huang et al.,[19] produced the highest imputation accuracy for that population. The genotypes of Pemberton et al.,[24] which update those reported by Conrad et al.,[25] were treated as true genotypes of the 513 markers for measurement of ε_{ij}. For each population, at each marker, the minor and major alleles were determined only with the use of the "true" genotype data from that population. If each allele had a frequency of 50%, then the minor allele was assigned at random.

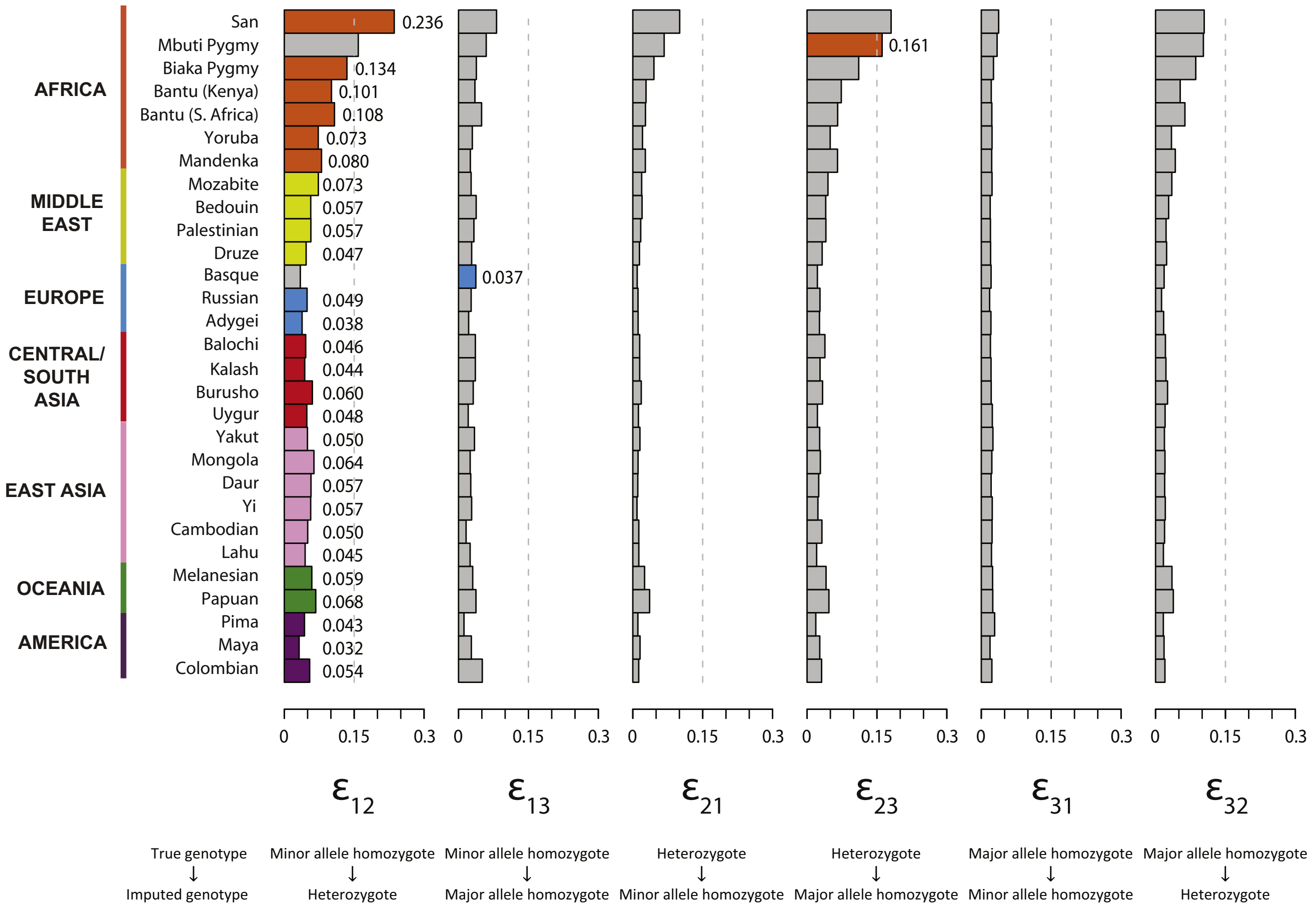

Figure 1. Genotype Misclassification Rates at Imputed Loci, in Each of 29 Populations
Each bar plot presents a particular error rate ε_{ij}, in which ε_{ij} represents the probability that genotype *i* is imputed as genotype *j* (1, minor allele homozygote; 2, heterozygote; 3, major allele homozygote). For each population, the greatest of the six error rates is shown in a color characteristic of the geographic region of the population. For convenience in interpreting the figure, the vertical dashed line indicates 15% error. The values plotted in the figure appear together with the overall imputation error rate in Table S1.

Treating the 426 individuals as unaffected, we classified 218,345 true genotypes (426 × 513, excluding missing data) by category, and for each population, we estimated ε_{12}, ε_{13}, ε_{21}, ε_{23}, ε_{31}, and ε_{32}. Each true genotype was categorized as follows: 1, minor allele homozygote; 2, heterozygote; 3, major allele homozygote. Considering all true genotypes in a population at all 513 markers, denote the number of true genotypes of types 1, 2, and 3 by n_1, n_2, and n_3, respectively. For each population, n_1, the smallest of the three quantities, was at least 70, so that at least 70 true genotypes were used in estimating each error parameter. For n_1, n_2, and n_3, the medians across populations were 411, 1967, and 3679, respectively.

To incorporate the uncertainty inherent in imputing a genotype, we obtained posterior probabilities of imputing types 1, 2, and 3. Considering the n_i genotypes of type *i*, denote the posterior probability that genotype ℓ was imputed to have type *j* by $q_{ij\ell}$. For each $i, j \in \{1, 2, 3\}$, $i \neq j$, we computed ε_{ij} for the population as $\sum_{\ell=1}^{n_i} q_{ij\ell}/n_i$. The "overall imputation error rate," a weighted average of the ε_{ij} that evaluates the total fraction of alleles imputed incorrectly, was calculated as $[(\varepsilon_{12}/2+\varepsilon_{13})n_1+(\varepsilon_{21}/2+\varepsilon_{23}/2)n_2+(\varepsilon_{31}+\varepsilon_{32}/2)n_3]/(n_1+n_2+n_3)$.

For each population, Figure 1 displays the estimated values of ε_{ij}. In most populations, the highest imputation error rate is ε_{12}, indicating that, conditional on true genotype, the highest-probability error is misclassification of a minor allele homozygote as a heterozygote. The next highest error rate is usually ε_{13} or ε_{23}, reflecting misclassification probabilities for minor allele homozygotes or heterozygotes, respectively, as major allele homozygotes. Misclassification probabilities for major allele homozygotes or heterozygotes as minor allele homozygotes (ε_{31} and ε_{21}, respectively) are generally low.

Treating the estimated values of ε_{ij} as parametric values, for each population, we evaluated the sample-size inflation factor *f* for various choices of the unknown MAF_{cases} and $\mathrm{MAF}_{controls}$. Because the difference $\delta = \mathrm{MAF}_{cases} - \mathrm{MAF}_{controls}$ can be viewed as a measure of the magnitude of the association at a disease locus, we reparametrized *f* in terms of δ and $\mathrm{MAF}_{controls}$. Thus, using observed levels of imputation error, we examined the properties of *f* across the range of possible frequencies for the disease allele in cases and

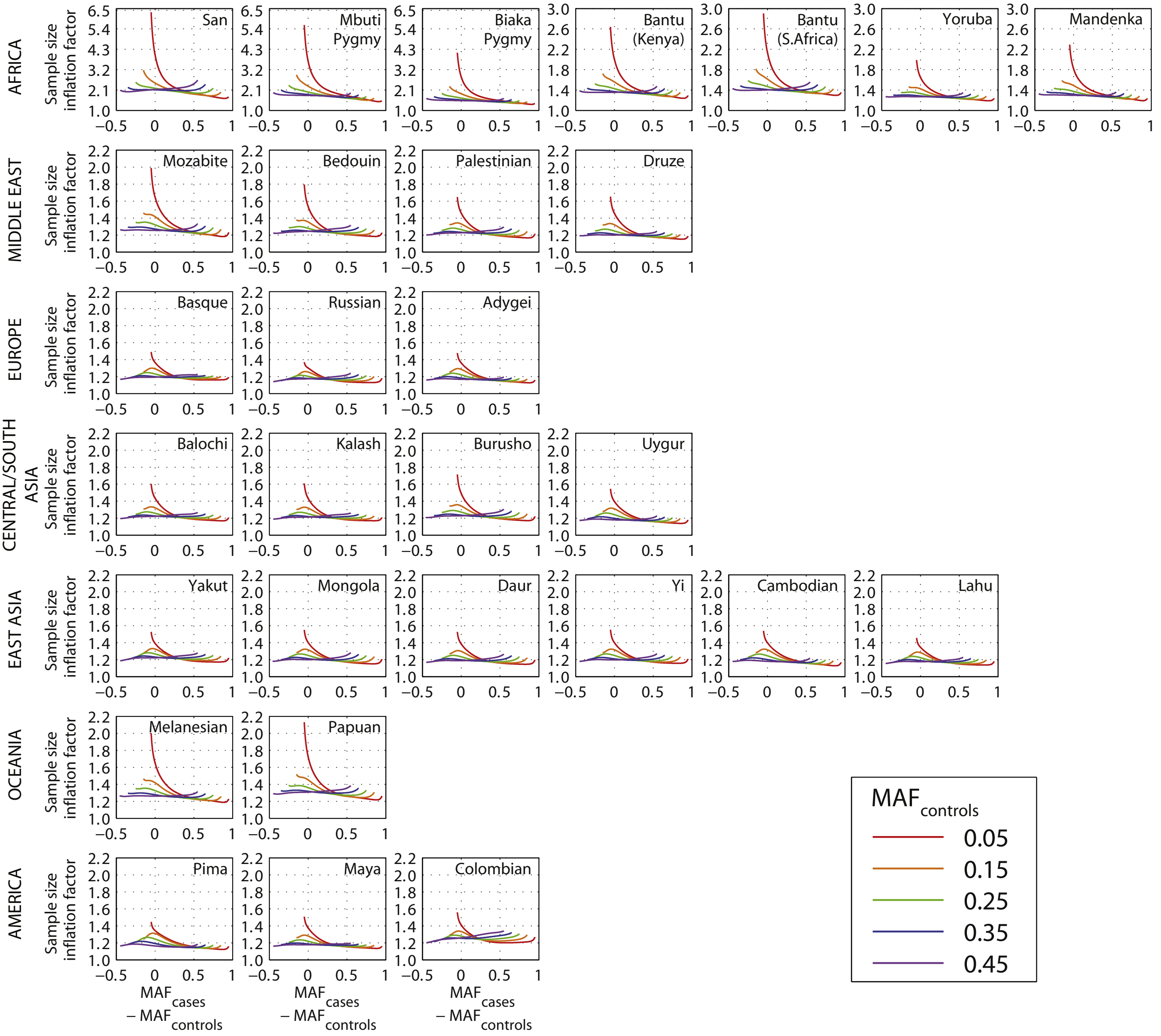

Figure 2. Sample-Size Inflation Factor *f* Required for Maintaining Statistical Power at Imputed Loci, as a Function of the True Difference in the Frequency of the Minor Allele between Cases and Controls

Each plot utilizes the estimated imputation error rates in Figure 1 for a specific population. For each population, the inflation factor is plotted for five choices of the true minor allele frequency in controls (0.05, 0.15, 0.25, 0.35, and 0.45). Note that $MAF_{controls}$ ranges from 0 to 0.5, whereas MAF_{cases}, representing the frequency in cases of the minor allele in controls, ranges from 0 to 1. We used a step size of 0.001 for MAF_{cases} and disregarded points with $MAF_{cases} = MAF_{controls}$.

controls (Figure 2). For most choices of the parameter values in most populations, the inflation factor *f* lies between 1.1 and 1.6. For most African populations, consistent with their higher imputation error rates, *f* is considerably greater than in other populations, ranging from 1.3 to 2.5 for most choices of the parameter values. The inflation factor is especially high in the San and Mbuti Pygmy populations, in which nearly all choices examined for δ and $MAF_{controls}$ produce $f \gtrsim 1.7$. Disease alleles are difficult to detect when |δ| is small, and Figure 2 demonstrates that for several populations, the sample-size inflation factor is greatest for small |δ|, particularly when the disease locus has a low minor allele frequency of $MAF_{controls} = 0.05$.

Because the parameters MAF_{cases} and $MAF_{controls}$ are unknown in actual association studies, for each population, conditional on the imputation error parameters ε_{ij}, we examined the minimal and maximal values of the sample-size inflation factor *f* across the range of possible values for MAF_{cases} and $MAF_{controls}$ (Figure 3). For most non-African populations, across most of the range of possible values for the minor allele frequency in controls, the minimal *f* is typically in the range of 1.1–1.2 and the maximum is typically in the range of 1.2–1.6, indicating that the extra sample size required for maintaining power is usually at least 10%–20% and at most 20%–60%. The maximal *f* is generally greater for low values of $MAF_{controls}$.

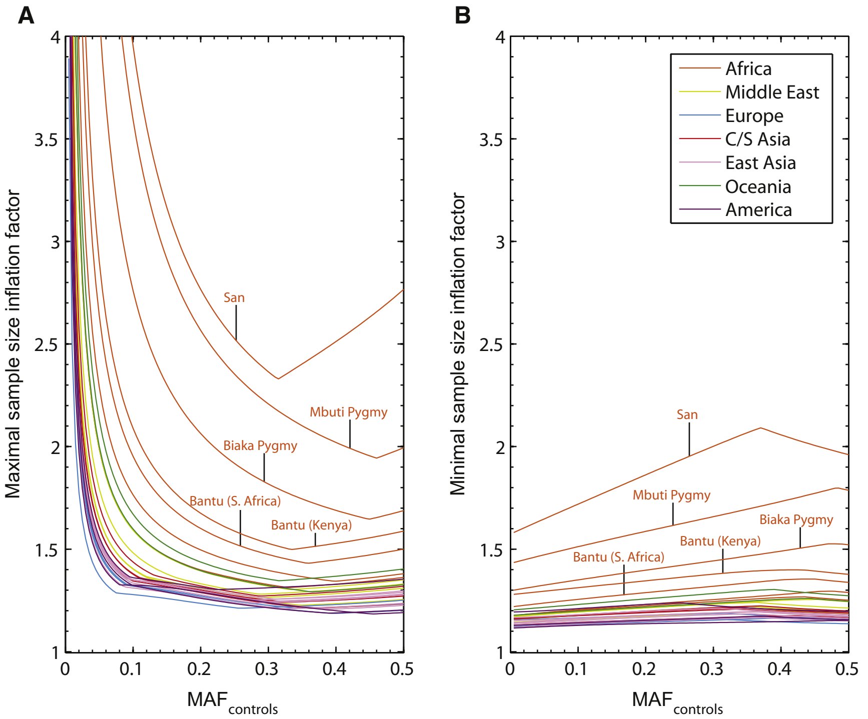

Figure 3. Maximal and Minimal Sample-Size Inflation Factors at Imputed Loci as Functions of the True Minor Allele Frequency in Controls, in Each of 29 Populations

For each value of $\mathrm{MAF}_{controls}$ from 0.005 to 0.5 with a step size of 0.005, the value plotted is the maximal or minimal value of the inflation factor f obtained across choices of MAF_{cases} ranging from 0 to 1 with a step size of 0.001 ($\mathrm{MAF}_{cases} \neq \mathrm{MAF}_{controls}$). Graphs for individual populations are color-coded by geographic region.
(A) Maximal sample-size inflation factor.
(B) Minimal sample-size inflation factor.

Upon examining the minimal and maximal sample-size inflation factor across the range of disease allele frequencies (Figure 3), we observe that the values are greatest in populations with the highest imputation error rates (Figure 1). Figure 4 quantifies this observation, illustrating the relationships with overall imputation error rate of the minimal and maximal values of f. A linear regression of the minimal sample-size inflation factor on overall imputation error rate when $\mathrm{MAF}_{controls}$ is fixed at 0.3, forced through the point at which no imputation errors occur and therefore no sample-size inflation occurs, provides a close fit for most populations, with the exceptions of the San and Mbuti Pygmy populations. The slope for this regression is 6.911, and the corresponding regression for the maximal sample-size inflation factor has a slope of 10.177. Excluding the San and Mbuti Pygmy populations, the slopes of the regressions for the minimal and maximal sample-size inflation factors decrease to 6.203 and 8.836, respectively (Figure S1, available online). The regression slopes generally lie between 5 and 13 when $\mathrm{MAF}_{controls}$ is fixed at various values across its range, with the San and Mbuti Pygmy populations either included or excluded (Figures S1 and S2). These values have the interpretation that each 1% increase in overall imputation error rate translates to an increase of ~5%–13% in the sample size required for maintaining power.

Our results have important implications for imputation studies. In the tag-SNP setting, for small values of x, a high LD level of $r^2 = 1 - x$ produces a relatively small sample-size inflation factor of $1/(1 - x) \approx 1 + x$, so that each 1% loss in the r^2 measure of LD leads to a ~1% gain in the required sample size. In the imputation setting, however, imputation accuracy of $1 - x$ produces a typical inflation factor of ~$1 + 5x$ to ~$1 + 13x$, so that each 1% loss in imputation accuracy leads to a ~5%–13% increase in the required sample size. As a result, even low levels of imputation error can have sizeable consequences. For example, measures that aim to assess genomic coverage for imputation methods might need to require stringent levels of imputation error in evaluating the proportion of the genome that is suited to imputation-based association mapping. Studies that aim to confirm associations at imputed markers in populations with lower imputation accuracy might inherently be disadvantaged for success in replication studies. In these various settings, careful assessment of appropriate sample sizes in power calculations will be essential for progress in imputation-based disease-gene identification. One key observation is that imputation error produces the greatest sample-size inflation for markers with low minor allele frequency ($\mathrm{MAF}_{controls} \leq 0.1$), and for such markers, the sample-size inflation for each 1% imputation error can be as high as ~15%–35% (Figures S1 and S2). As GWA efforts begin to focus on the impact of rare alleles on complex diseases, the potentially serious effects of imputation error for detecting such alleles will be a central consideration for forthcoming studies. For such studies, it will be informative to examine values of the imputation error parameters ε_{ij} evaluated specifically from rare alleles.

We note that the linear dependence of the minimal and maximal sample-size inflation factor on overall imputation error rate, as illustrated in Figure 4, is only approximate. This approximate linear relationship arises because the overall imputation error rate is a composite parameter dependent on the six underlying ε_{ij} parameters, each of which affects the inflation factor in an approximately linear manner. On the basis of a first-order Taylor series expansion for f, for each i and j, Kang et al.[22] derived cost functions C_{ij}, so that if all error parameters except ε_{ij} are set to zero and ε_{ij} is small, then the sample-size inflation factor is approximately $1 + C_{ij}\varepsilon_{ij}$. These linear approximations accurately reflect the sample-size inflation factor in most populations except at the lowest values of MAF_{cases} and

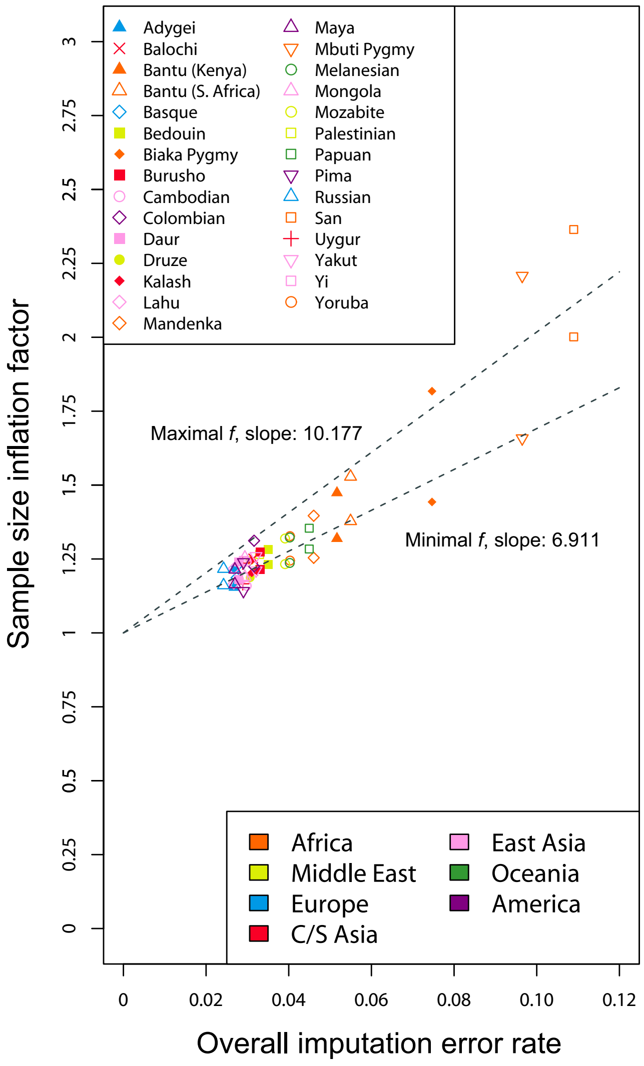

Figure 4. Maximal and Minimal Sample-Size Inflation Factors as Functions of the Overall Imputation Error Rate, for an Imputed Disease Locus with a True Minor Allele Frequency of 0.3 in Controls

Populations are color-coded by geographic region, and two data points appear for each population: a maximum and a minimum. Best-fit linear-regression lines for the maxima and minima, forced through the point (0,1), indicate the increase in the inflation factor with increasing imputation error rate. For example, the lines indicate that in most populations, at $MAF_{controls} = 0.3$, imputation error rates of 2%–6% correspond to sample-size inflation factors of ~14%–53%, and each additional 1% increase in imputation error corresponds to a ~7%–10% increase in the inflation factor.

$MAF_{controls}$ (results not shown) and suggest that in general, the greatest cost is incurred from errors in imputing minor allele homozygotes as major allele homozygotes (Figure 5). It is noteworthy that the linear regressions in Figure 4 provide the poorest underestimates in the San population, for which the parameter ε_{13} for the most costly type of error is high and for which the pattern of errors differs somewhat from the corresponding patterns in the other populations (Figure 1).

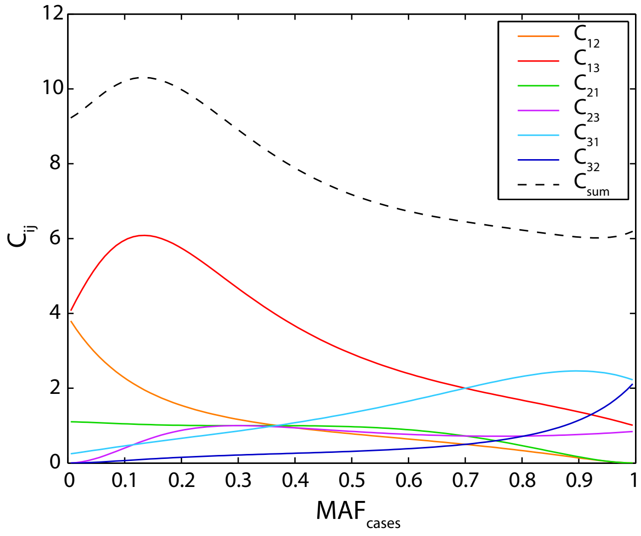

Figure 5. Cost Coefficients as Functions of MAF_{cases} for the Fixed Value $MAF_{controls} = 0.3$

The coefficient C_{ij} provides an approximation for the relative magnitude of the sample-size inflation that is due to the error parameter ε_{ij}. Thus, a small increase of x in the imputation error parameter ε_{ij} adds approximately $C_{ij}x$ to the sample-size inflation factor. The sum of the six cost coefficients, C_{sum}, has the interpretation that $C_{sum}x$ is added to the sample-size inflation factor when all six of the ε_{ij} are simultaneously set to x. Each of the cost coefficients was evaluated for values of MAF_{cases} from 0.005 to 0.995 at intervals of 0.01.

Although an increased sample size provides one approach to maintaining power in an imputation-based study, an alternative strategy is to decrease imputation error instead. Reductions in imputation error can be achieved through a combination of algorithmic advances and optimal choices of imputation algorithms,[13,15] improvements in usage of existing reference panels,[19,26] and expanded marker density and sample inclusion in these panels.[18,27] A fourth approach involves incorporating information on relatives of study subjects for the improvement of phase estimates at measured markers. Although this approach will not eliminate errors owing to incorrect imputation conditional on correctly estimated phase, it will reduce imputation errors that arise from incorrect phase estimation.

For populations with relatively little imputation error in which large samples are easily obtained, the required sample-size increase produced by imputation error might not pose a significant obstacle for GWA studies. In other populations in which subject recruitment is difficult and the sample-size inflation required for maintaining power is extreme, reduction of imputation error might be more feasible than an increase in sample size. As GWA studies begin diversifying to incorporate additional populations beyond the populations of European origin that have been typical of most investigations to date,[28] it will be important to evaluate the relative merits of the various approaches for overcoming the consequences of imputation error to improve the potential of imputation-based association studies.

Supplemental Data

Supplemental Data include two figures and one table and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

We thank M. Boehnke, Y. Li, and two reviewers for helpful comments. This work was supported by National Institutes of Health grants R01 GM081441 and R01 HL090564, by a Burroughs Wellcome Fund Career Award in the Biomedical Sciences, and by an Alfred P. Sloan Research Fellowship.

Received: July 13, 2009
Revised: September 15, 2009
Accepted: September 24, 2009
Published online: October 22, 2009

References

1. Li, Y., Ding, J., and Abecasis, G.R. (2006). Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. Am. J. Hum. Genet. *79*, S2290.
2. Nicolae, D.L. (2006). Testing untyped alleles (TUNA) — applications to genome-wide association studies. Genet. Epidemiol. *30*, 718–727.
3. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. *39*, 906–913.
4. Servin, B., and Stephens, M. (2007). Imputation-based analysis of association studies: Candidate regions and quantitative traits. PLoS Genet. *3*, 1296–1308.
5. Browning, S.R. (2008). Missing data imputation and haplotype phase inference for genome-wide association studies. Hum. Genet. *124*, 439–450.
6. Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U., et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science *316*, 1341–1345.
7. 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.
8. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., et al. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. *40*, 955–962.
9. Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., and Diabetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat. Genet. *5*, 638–645.
10. Willer, C.J., Sanna, S., Jackson, A.U., Scuteri, A., Bonnycastle, L.L., Clarke, R., Heath, S.C., Timpson, N.J., Najjar, S.S., Stringham, H.M., et al. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat. Genet. *40*, 161–169.
11. Yu, Z., and Schaid, D.J. (2007). Methods to impute missing genotypes for population data. Hum. Genet. *122*, 495–504.
12. Guan, Y., and Stephens, M. (2008). Practical issues in imputation-based association mapping. PLoS Genet. *4*, e1000279.
13. Pei, Y.-F., Li, J., Zhang, L., Papasian, C.J., and Deng, H.-W. (2008). Analyses and comparison of accuracy of different genotype imputation methods. PLoS ONE *3*, e3551.
14. Zhao, Z., Timofeev, N., Hartley, S.W., Chui, D.H.K., Fucharoen, S., Perls, T.T., Steinberg, M.H., Baldwin, C.T., and Sebastiani, P. (2008). Imputation of missing genotypes: An empirical evaluation of IMPUTE. BMC Genet. *9*, 85.
15. Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M., and Franke, A. (2009). A comprehensive evaluation of SNP genotype imputation. Hum. Genet. *125*, 163–171.
16. Pritchard, J.K., and Przeworski, M. (2001). Linkage disequilibrium in humans: Models and data. Am. J. Hum. Genet. *69*, 1–14.
17. de Bakker, P.I.W., Ferreira, M.A.R., Jia, X., Neale, B.M., Raychaudhuri, S., and Voight, B.F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum. Mol. Genet. *17*, R122–R128.
18. Browning, B.L., and Browning, S.R. (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.
19. Huang, L., Li, Y., Singleton, A.B., Hardy, J.A., Abecasis, G., Rosenberg, N.A., and Scheet, P. (2009). Genotype imputation accuracy across worldwide human populations. Am. J. Hum. Genet. *84*, 235–250.
20. Lettre, G., Jackson, A.U., Gieger, C., Schumacher, F.R., Berndt, S.I., Sanna, S., Eyheramendy, S., Voight, B.F., Butler, J.L., Guiducci, C., et al. (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. Nat. Genet. *40*, 584–591.
21. Gordon, D., Finch, S.J., Nothnagel, M., and Ott, J. (2002). Power and sample size calculations for case-control genetic association tests when errors are present: Application to single nucleotide polymorphisms. Hum. Hered. *54*, 22–33.
22. Kang, S.J., Gordon, D., and Finch, S.J. (2004). What SNP genotyping errors are most costly for genetic association studies? Genet. Epidemiol. *26*, 132–141.
23. Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.-C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., et al. (2008). Genotype, haplotype, and copy-number variation in worldwide human populations. Nature *451*, 998–1003.
24. Pemberton, T.J., Jakobsson, M., Conrad, D.F., Coop, G., Wall, J.D., Pritchard, J.K., Patel, P.I., and Rosenberg, N.A. (2008). Using population mixtures to optimize the utility of genomic databases: Linkage disequilibrium and association study design in India. Ann. Hum. Genet. *72*, 535–546.
25. Conrad, D.F., Jakobsson, M., Coop, G., Wen, X., Wall, J.D., Rosenberg, N.A., and Pritchard, J.K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat. Genet. *38*, 1251–1260.
26. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.
27. Becker, T., Flaquer, A., Brockschmidt, F.F., Herold, C., and Steffens, M. (2009). Evaluation of potential power gain with imputed genotypes in genome-wide association studies. Hum. Hered. *68*, 23–34.
28. Cooper, R.S., Tayo, B., and Zhu, X. (2008). Genome-wide association studies: Implications for multiethnic samples. Hum. Mol. Genet. *17*, R151–R155.