Evaluating the Effects of Imputation on the Power, Coverage, and Cost Efficiency of Genome-wide SNP Platforms

Carl A. Anderson,[1,*] Fredrik H. Pettersson,[1] Jeffrey C. Barrett,[1] Joanna J. Zhuang,[1] Jiannis Ragoussis,[1] Lon R. Cardon,[2] and Andrew P. Morris[1]

Genotype imputation is potentially a zero-cost method for bridging gaps in coverage and power between genotyping platforms. Here, we quantify these gains in power and coverage by using 1,376 population controls that are from the 1958 British Birth Cohort and were genotyped by the Wellcome Trust Case-Control Consortium with the Illumina HumanHap 550 and Affymetrix SNP Array 5.0 platforms. Approximately 50% of genotypes at single-nucleotide polymorphisms (SNPs) exclusively on the HumanHap 550 can be accurately imputed from direct genotypes on the SNP Array 5.0 or Illumina HumanHap 300. This roughly halves differences in coverage and power between the platforms. When the relative cost of currently available genome-wide SNP platforms is accounted for, and finances are limited but sample size is not, the highest-powered strategy in European populations is to genotype a larger number of individuals with the HumanHap 300 platform and carry out imputation. Platforms consisting of around 1 million SNPs offer poor cost efficiency for SNP association in European populations.

The advent of cost-effective, high-throughput genotyping technologies has greatly aided the identification of genetic variants underlying human disease.[1–4] Most genome-wide association (GWA) studies use mass-produced genotype chips designed to capture a given proportion of genetic variation genome wide. The exact proportion of captured variation differs among platforms and populations depending on the number of single-nucleotide polymorphisms (SNPs) and their selection criteria.[5] If sample size is limited but finances are not, individuals should be genotyped with the platform that provides the most genomic coverage. When sample size is unlimited but finances are restricted, a choice must be made between genotyping the maximum number of individuals or the maximum number of SNPs.

Imputation can potentially bridge the gap in coverage between genome-wide SNP platforms. Here, direct genotype data are combined with population haplotype and historical recombination information to predict an individual's genotype at ungenotyped SNPs. If all SNPs featured exclusively on a given chip can be accurately imputed on the basis of genotype data from a platform of lower density (and cost), one should always choose to genotype additional individuals in preference to additional SNPs. In practice, not all of these SNPs are going to be imputed accurately. Therefore, to ascertain whether the number of SNPs or the number of individuals should be maximized in the design of a GWA, two further questions must be answered: (1) What proportion of SNPs found exclusively on a dense GWA platform can be accurately imputed from genotype data on less dense products? (2) What effect in terms of coverage and power do these imputed SNPs have on genome-wide association analysis and our choice of platform?

To answer these questions, we used empirical data from the 1958 British Birth Cohort (58C)[6] genotyped with both the HumanHap 550 platform and SNP Array 5.0 by the Wellcome Trust Case-Control Consortium (WTCCC).[4] For a given platform, individuals with a genome-wide genotype missingness $> 5\%$ were removed in addition to related, duplicated, and individuals of non-European descent identified during the original WTCCC project.[4] Only the 1,376 individuals passing quality control (QC) on both SNP Array 5.0 and HumanHap 550 platforms remained under study. For both platforms, SNPs with a genotype call rate $< 95\%$, or $< 99\%$ for SNPs with a minor-allele frequency (MAF) $< 5\%$, were removed in addition to SNPs with a Hardy-Weinberg exact p value $< 5.7 \times 10^{-7}$. Of the 555,352 SNPs on the HumanHap 550 platform, 529,167 (95.3%) passed QC. Of the 500,568 SNPs on the SNP Array 5.0, 459,450 (91.7%) passed QC. Direct genotyping was not carried out with the HumanHap 300 platform, so we use the 313,504 SNPs (98.74% of all HumanHap 300 SNPs) that are also featured on the HumanHap 550 as a proxy (HumanHap 300*).

We excluded 22,930 SNP Array 5.0 SNPs and 22,416 HumanHap 300 SNPs that were monomorphic in the CEU HapMap panel. Some of these SNPs were polymorphic in the 58C, and this highlights one disadvantage of basing imputations on the small number of individuals in the HapMap—variation at rare SNPs (MAF $\leq 2\%$) that are monomorphic in the HapMap will be lost (Figure 1). Phase III of the HapMap project is currently ongoing and aims at genotyping the ~1.7 million SNPs that are featured on the HumanHap 1M or Affymetrix SNP Array 6.0. Approximately twice the number of samples currently in the HapMap will be genotyped. This improved resource should enable rare SNPs to be imputed more accurately. In the

[1]The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, OX3 7BN Oxford, UK; [2]Fred Hutchinson Cancer Research Center, Mail Stop C3-168, P.O. Box 19024, Seattle, WA 98109-1024, USA
*Correspondence: carl.anderson@well.ox.ac.uk
DOI 10.1016/j.ajhg.2008.06.008.

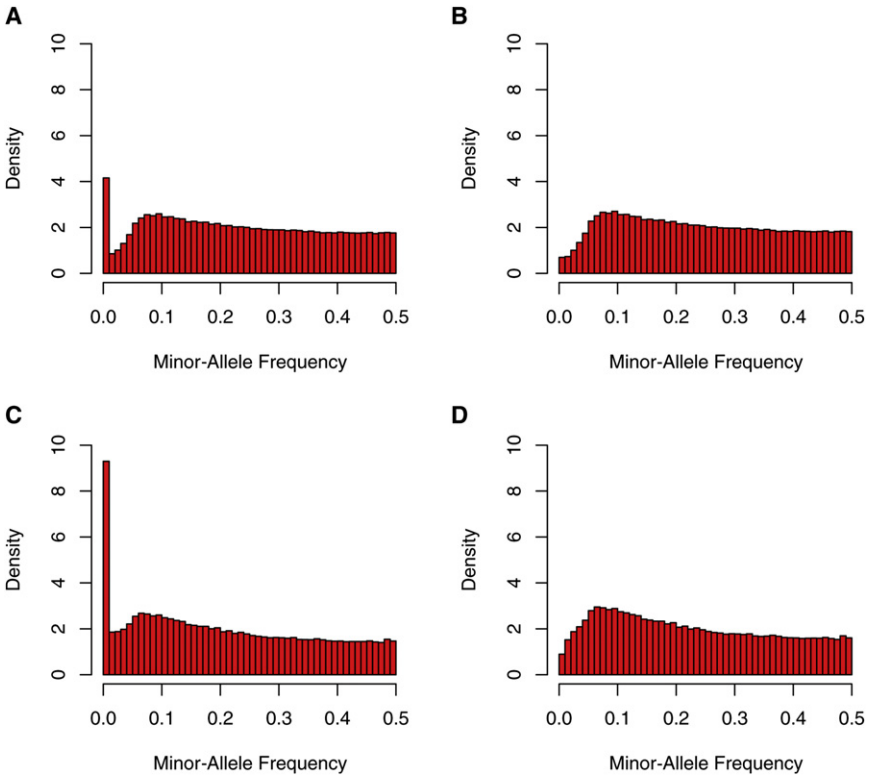

Figure 1. Minor-Allele Frequency of SNPs Directly Genotyped in 1,376 Samples from the 58C

(A) Minor-allele frequency for the 450,769 SNPs that are featured on the HumanHap 550 but not the Affymetrix SNP Array 5.0 and are also polymorphic in the 58C.

(B) Minor-allele frequency for the subset of 427,839 SNPs from (A) that are also polymorphic in the CEU HapMap data.

(C) Minor-allele frequency for the 215,998 that are featured on the HumanHap 550 but not the Illumina HumanHap 300* and are also polymorphic in the 58C.

(D) Minor-allele frequency for the subset of 203,860 SNPs from (C) that are also polymorphic in the CEU HapMap data. Basing imputations on haplotype data from the HapMap causes variation at rare SNPs (MAF $\leq$ 0.02) to be lost.

present study, imputations were based on CEU haplotype information for all SNPs found in Phase II HapMap release 21 - NCBI Build 35 (dbSNP build 125) and data detailing local recombination rates.[7,8] These were combined with direct genotyping data from either the SNP Array 5.0 or HumanHap 300*. For each of these platforms, we imputed all polymorphic HapMap SNPs (excluding those for which we had direct genotype data), although only SNPs that are featured exclusively on the HumanHap 550 were taken forward for analysis. We could not use all HapMap SNPs in our analysis because we only had direct genotypes available for comparison for those SNPs on the HumanHap 550 or SNP Array 5.0. Imputations were carried out with IMPUTE,[9] which outputs the posterior probability that an individual is each of three genotype classes (AA, AB, BB) per SNP. For a given individual, any genotype with a posterior probability $\geq$ 0.9 was called, and genotypes at SNPs with a maximum posterior probability below this threshold were classified as missing. Typically, when analyzing imputed genotype data, one should average over the distribution of genotype probabilities and carry out a weighted logistic regression analysis. However, here we wished to directly compare the direct and imputed genotypes and so created an "imputed genotype." In total, we contrasted imputed genotypes at 427,838 SNPs on the basis of the SNP Array 5.0 and 203,859 SNPs on the basis of the HumanHap 300* to direct genotypes from the lHumanHap 550 platform.

Before comparing the direct and imputed genotypes, we removed SNPs that were likely to be imputed inaccurately. These include SNPs that (1) have swapped strands between HapMap release21 and release22; (2) are removed from the latest versions of genome-wide SNP chips; (3) have more than 5% missing data in the HapMap; (4) have more than 5% discordant genotypes between original HapMap genotyping and the regenotyping of the CEU HapMap samples using genome-wide SNP chips; and (5) show significant differences in frequency when called with the BRLMM and CHIAMO++[4] genotype-calling algorithms. In total, 45,692 and 31,953 SNPs on the HumanHap 550 were removed from the imputations based on SNP Array 5.0 and HumanHap 300, respectively.

We applied partial least-squares projection to latent-structures discriminant analysis (PLS-DA) to detect and quantify systematic differences between direct and imputed genotypes. PLS-DA is a multivariate regression method that relates a data matrix ($\mathbf{X}$) to a singular ($\mathbf{y}$) or multiple ($\mathbf{Y}$) response variables. In the current study, $\mathbf{X}$ is a single matrix containing 1,376 × 2 rows (each individual is represented twice, once with a direct genotype vector and once with an imputed genotype vector) and n columns, where n is the number of SNPs under comparison. $\mathbf{Y}$ is a vector classifying each individual genotype vector (row) as either direct or imputed genotypes To get a numerical estimate of how strongly a model based on the direct and imputed genotypes can discriminate if a particular genotype vector is from a direct or imputed source, we applied a cross-validation approach. Here, supervised models are built with six-sevenths of the genotype vectors (both direct and imputed). Each model is used to predict the origin of the remaining one-seventh of genotype vectors. For this "test set" of genotype vectors, the prediction model uses only the genotypes to predict status (i.e., the known status of the genotype vector is ignored). This procedure is repeated seven times until the status of each genotype

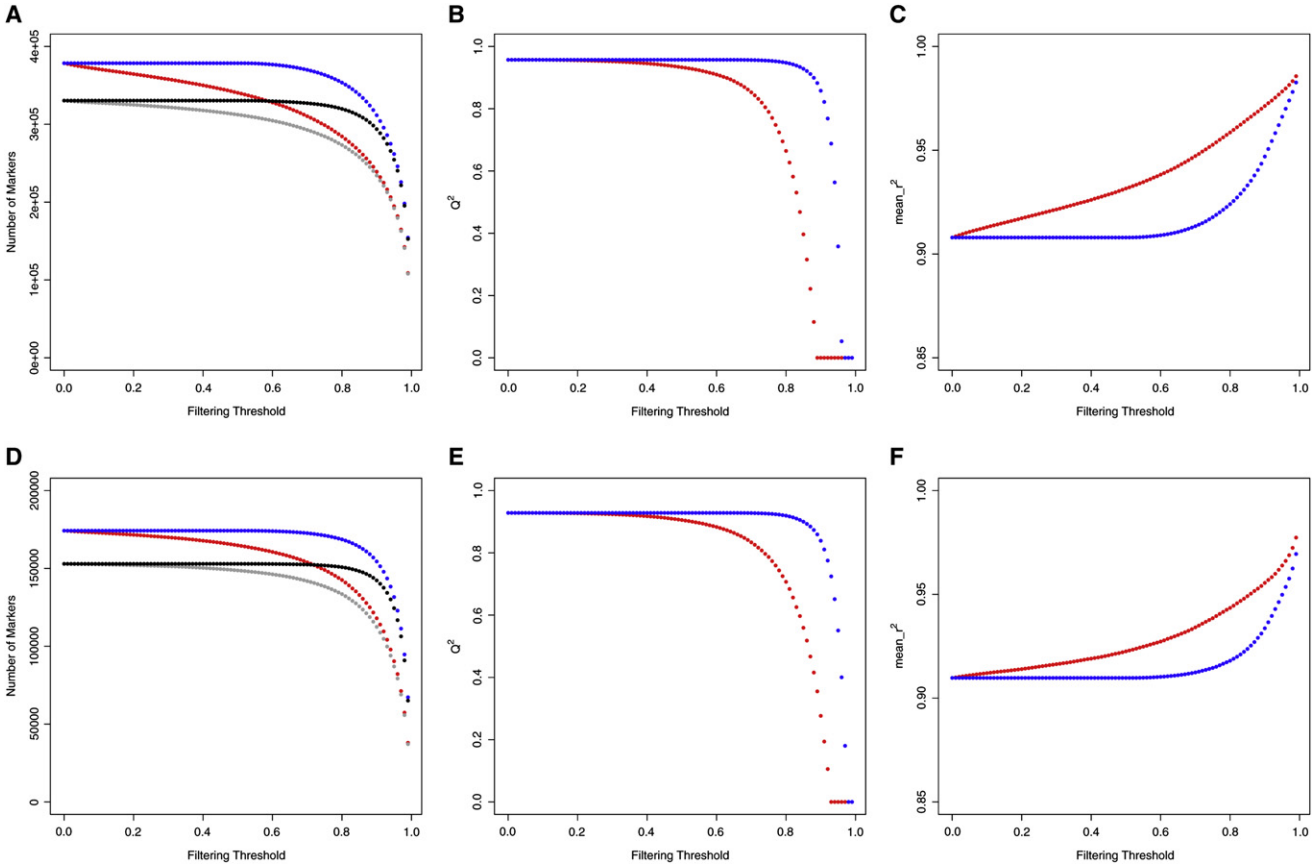

Figure 2. Assessment of Imputed-Genotype Filtering Criteria

Assessment of filtering criteria for the Illumina HumanHap 550 genotypes based on Affymetrix SNP Array 5.0 (A–C) and Illumina HumanHap300* (D–F) genotype data.

(A and D) The number of SNPs passing filter thresholds based on per-SNP measures of mean maximum posterior probability (blue) or genotype call rate (red). The number of these SNPs with an $r^2 \geq 0.8$ between direct and imputed genotype calls is shown after the removal of SNPs not passing filtering thresholds based on per-SNP measures of mean maximum posterior probability (dark gray) and genotype call rate (light gray).

(B and E) The PLS-DA Q^2 value after the removal of SNPs not passing filtering thresholds based on per-SNP measures of mean maximum posterior probability (blue) and genotype call rate (red). A Q^2 value of 1 indicates that the current PLS model can perfectly predict whether a given genotype vector is of direct or imputed origin. A Q^2 of 0 indicates that the model has no power to predict the genotype's origin.

(C and F) Mean r^2 between direct and imputed genotypes after the removal of SNPs not passing filtering thresholds based on per-SNP measures of mean maximum posterior probability (blue) and genotype call rate (red).

vector has been predicted once. Subsequently, a cross-validation score (Q^2) is calculated from the difference between the observed and predicted genotype status values. Q^2 reflects the mean accuracy of the models for predicting the status of the genotype vectors.[10] Therefore, a Q^2 of 1 indicates that many systematic differences exist between direct and imputed genotypes, and a Q^2 of 0 indicates that no such differences exist.

Filtering imputed data based on genotype call rate and/or mean maximum posterior probability can remove many poorly imputed genotypes. For both metrics, we incremented the filtering threshold from 0 to 1 in steps of 0.01 to ascertain which filtering criterion most efficiently eliminated differences between direct and imputed genotypes. A PLS-DA model was fitted to each data set, and Q^2 was estimated through cross-validation to assess the level of dispersion. For each filtering criterion, the number of remaining SNPs and the mean r^2 between the remaining direct and imputed genotypes were calculated. Filtering imputed SNPs based on genotype call rate was more efficient than mean maximum posterior probability (Figure 2). Using SNP Array 5.0 genotypes, we were able to accurately impute 245,430 of the attempted 427,838 HumanHap 550 SNPs after removing all SNPs with an imputed genotype call rate ≤ 0.89. For imputations based on HumanHap 300* genotypes, we were able to accurately impute 104,180 of the attempted 203,859 HumanHap 550 SNPs after removing all those with a genotype call rate ≤ 0.93. The applied filtering criteria represent the most efficient means of removing systematic differences between direct and imputed genotypes in our data. When we repeated the analysis by including the SNPs that we believed, a priori, were likely to be imputed badly, more stringent thresholds were required and fewer SNPs were successfully imputed.

Table 1. Estimates of Genomic Coverage for Currently Available Genome-wide SNP Platforms Alone and after Imputation

	Percentage of Genomic Coverage at $r^2 \geq 0.8$	Percentage of Genomic Coverage at $r^2 = 1$
Affymetrix SNP Array 5.0	65	43
Affymetrix SNP Array 5.0 plus imputed SNPs	73	54
Affymetrix SNP Array 6.0	80	59
Illumina HumanHap 300	77	42
Illumina HumanHap 300 plus imputed SNPs	81	50
Illumina HumanHap 550	87	57
Illumina HumanHap 1M	91	68

Estimates evaluated with Phase II HapMap data from the CEU population. Coverage estimates for Illumina HumanHap 1M and Affymetrix SNP-array-6.0 are likely to be biased downward because the genotypes at approximately 10% of the SNPs on each platform are not currently publicly available for the CEU HapMap individuals. Where imputations are included, all SNPs passing imputation-filter thresholds and with an $r^2 \geq 0.8$ between known and imputed genotypes are included along with the SNPs on the genome-wide SNP chip.

The certainty that an imputation algorithm has for a given imputed genotype does not always perfectly reflect the accuracy of imputation. Estimates of coverage and power would be biased if we assumed all SNPs passing the imputed genotype call-rate filters were accurate, so only those with an $r^2 \geq 0.8$ between known and imputed genotypes were taken forward for further analysis (239,931 and 99,831 SNPs imputed from the SNP Array 5.0 and HumanHap 300, respectively). Coverage was estimated with the formula outlined by Barrett and Cardon,[5] and results showed that approximately half the difference in coverage between the SNP Array 5.0 or HumanHap 300 and the HumanHap 550 could be recovered through imputation (Table 1).

We carried out simulations with the R statistical package[11] to quantify the power to detect association with SNPs present on the (1) SNP Array 5.0; (2) HumanHap 300; (3) SNP Array 5.0 plus successful imputations; (4) HumanHap 300 plus successful imputations; (5) Affymetrix SNP-array-6.0, and (6) Illumina HumanHap 1M. For each polymorphic CEU HapMap SNP (release 21), we calculated the maximum r^2, in the CEU HapMap data, between it and the SNPs on each of the above genotyping platforms by using the CEU HapMap data. For SNPs that are featured on the given SNP chip, maximum r^2 is therefore 1. We assessed the power of each SNP set to detect associations to SNPs of varying allelic odds ratio (1.0–2.0, 0.01 increments) and risk-increasing allele frequency (RAF) ($0.05 \leq RAF < 0.10$, $0.10 \leq RAF < 0.20$, $0.20 \leq RAF \leq 0.50$). To obtain empirical distributions of RAF and maximum r^2, we selected at random 10,000 HapMap SNPs within each of the RAF ranges. If the causal variant is typed and n individuals are needed to obtain a given power, then n/r^2 individuals are needed to obtain the same power if a tagSNP is typed where the correlation between the causal SNP and the tagSNP is r^2. Therefore, for a given SNP, the maximum r^2 can be used to weight the sample size, and hence the power to detect an association can be calculated for each of the SNP platforms under study. Power was calculated with an assumed type I error rate of 10^{-5} and a multiplicative disease model. For each RAF range and allelic odds ratio, the mean power across all 10,000 SNPs was calculated to provide an unbiased measure of power (see Appendix 1).

Initially, fixed baseline sample sizes of 1,000, 2,000, and 5,000 cases (and an equal number of controls) were simulated across all platforms. The successfully imputed SNPs significantly improved the power of the SNP Array 5.0 and HumanHap 300 to detect association to rare SNPs ($0.05 \leq MAF < 0.1$) of reasonable effect (1.6 – 2.0), yielding similar power as the SNP-array-6.0 and each other (Figure 3). For more common SNPs (or smaller effect sizes), the differences between the platforms are less pronounced, so the effect of imputation was reduced.

To reflect the cost efficiencies of the various platforms, we carried out simulations where the baseline sample size for a given platform was multiplied by the ratio in price per individual sample between that platform and the HumanHap 550 (sample-size ratios: SNP Array 5.0 = 1.22; HumanHap 300 = 1.32; HumanHap 550 = 1; SNP-array-6.0 = 0.99; HumanHap 1M = 0.57). These price-per-sample ratios were calculated on the basis of indicative UK prices (including work costs) communicated to us directly from Affymetrix and Illumina and are included here as a guide only. Strikingly, under all simulated scenarios, the HumanHap 300 with imputation provides the most power (Figure 4). This is because it is the cheapest of the GWA panels but provides, with successful imputations, good genomic coverage (81%) of common variation ($MAF \geq 0.05$) at $r^2 \geq 0.8$ in populations of European descent. Even without imputation, the HumanHap 300 platform appears to be the platform of choice if sample size is unlimited but finances are not largely due to a very competitive pricing. In terms of cost efficiency for power to detect associations in samples of European descent, the HumanHap 1M platform is of low value. However, ~10% of SNPs on this (and the SNP-array-6.0) platform are not typed in HapMap samples and are not included in our analyses. Furthermore, the HumanHap 1M has many tagSNPs for African populations, which probably offer little in terms of power and coverage in European populations. Because of a lack of empirical genotype data, we are unable to repeat our analyses for non-European populations.

Genome-wide SNP chips no longer focus solely on capturing variation at SNPs in populations of European descent and often include tagSNPs for other populations or tags to capture copy-number variation (CNV). Indeed, the HumanHap 550 has now been replaced by the HumanHap 610, which adds ~60,000 tags for CNVs. The implications of this on our work are minimal because the vast majority of these CNV tags are not SNPs and therefore have little effect on the coverage of, and power to detect association to, variation in SNPs. However, as the features on the various platforms become more varied, choosing

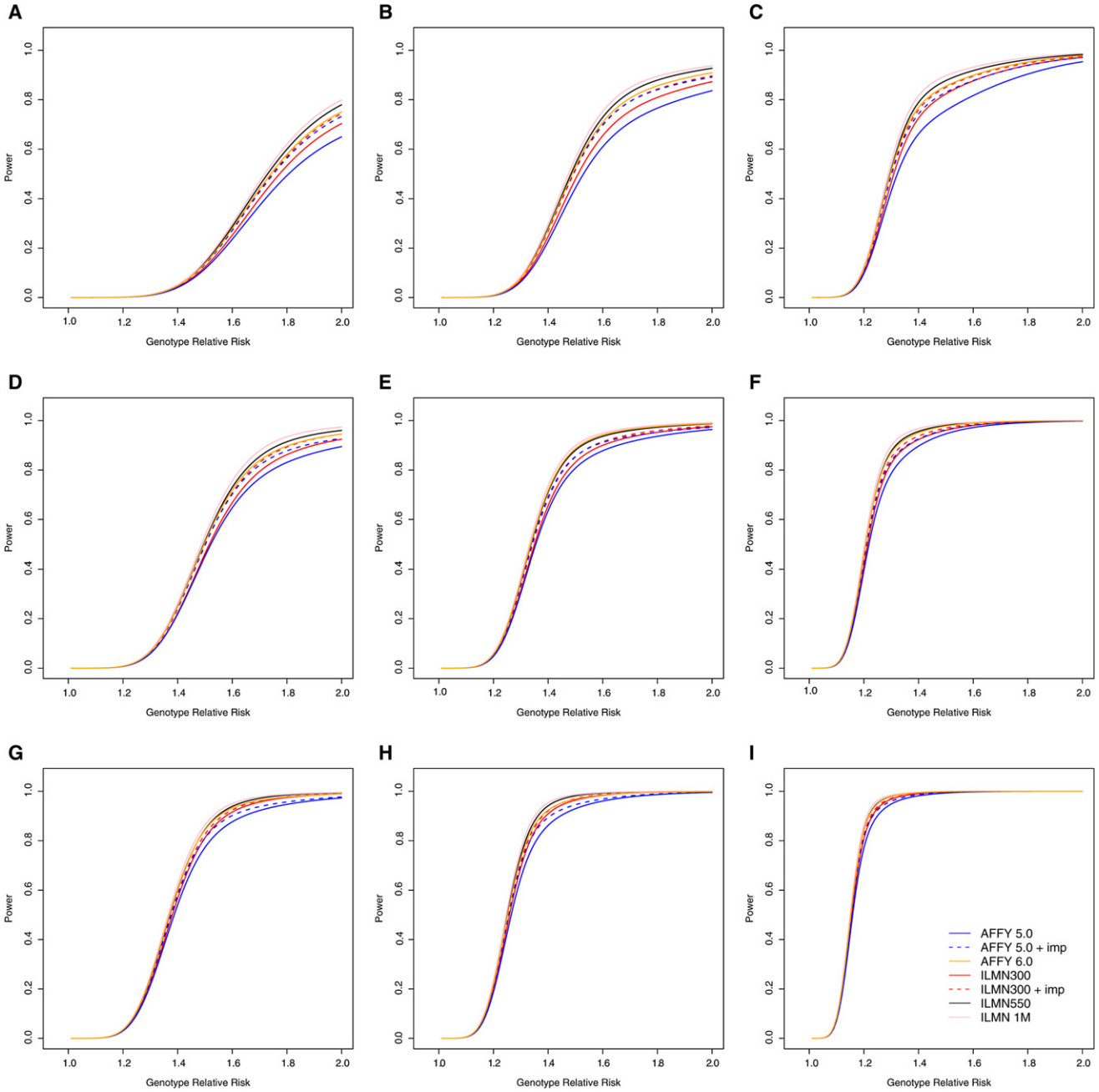

Figure 3. Mean power to Detect Association to a Disease with a Fixed Baseline Sample Size

Mean power to detect association ($\alpha = 10^{-5}$) to a disease with a population prevalence of 0.0001 and a fixed baseline sample size across different genome-wide platforms (simulated under varying risk allele frequency [RAF] and sample size).

RAF ranges are as follows: (A–C) $0.05 \leq \text{RAF} < 0.10$; (D–F) $0.10 \leq \text{RAF} < 0.20$; (G and H) $0.20 \leq \text{RAF} \leq 0.50$. Cases and controls are as follows: (A, D, and G) 1,000 cases, 1,000 controls; (B, E, and F) 2,000 cases, 2,000 controls; (C, F, and I) 5,000 cases, 5,000 controls. Mean power was calculated after 10,000 simulations where sample size per simulation for each SNP set was weighted by the maximum r^2 between a randomly selected HapMap SNP (satisfying RAF constraints) and the SNPs on the given genotyping platform (with HapMap release 21 CEU data).

a platform for a particular study becomes more difficult and depends largely on the study population and the believed underlying genetics of the trait. For example if rare CNVs are thought to play a role, then a GWA platform featuring CNV tags or other complementary platforms should be considered.[12] If the sample consists of individuals of non-European descent, then it is wise to obtain a platform that covers variation in that population. If a platform that does not have these added features is chosen, genotype imputation is unlikely to "recapture" this variation.

Without empirical genotyping data from other populations, it is difficult to assess how transferable our results

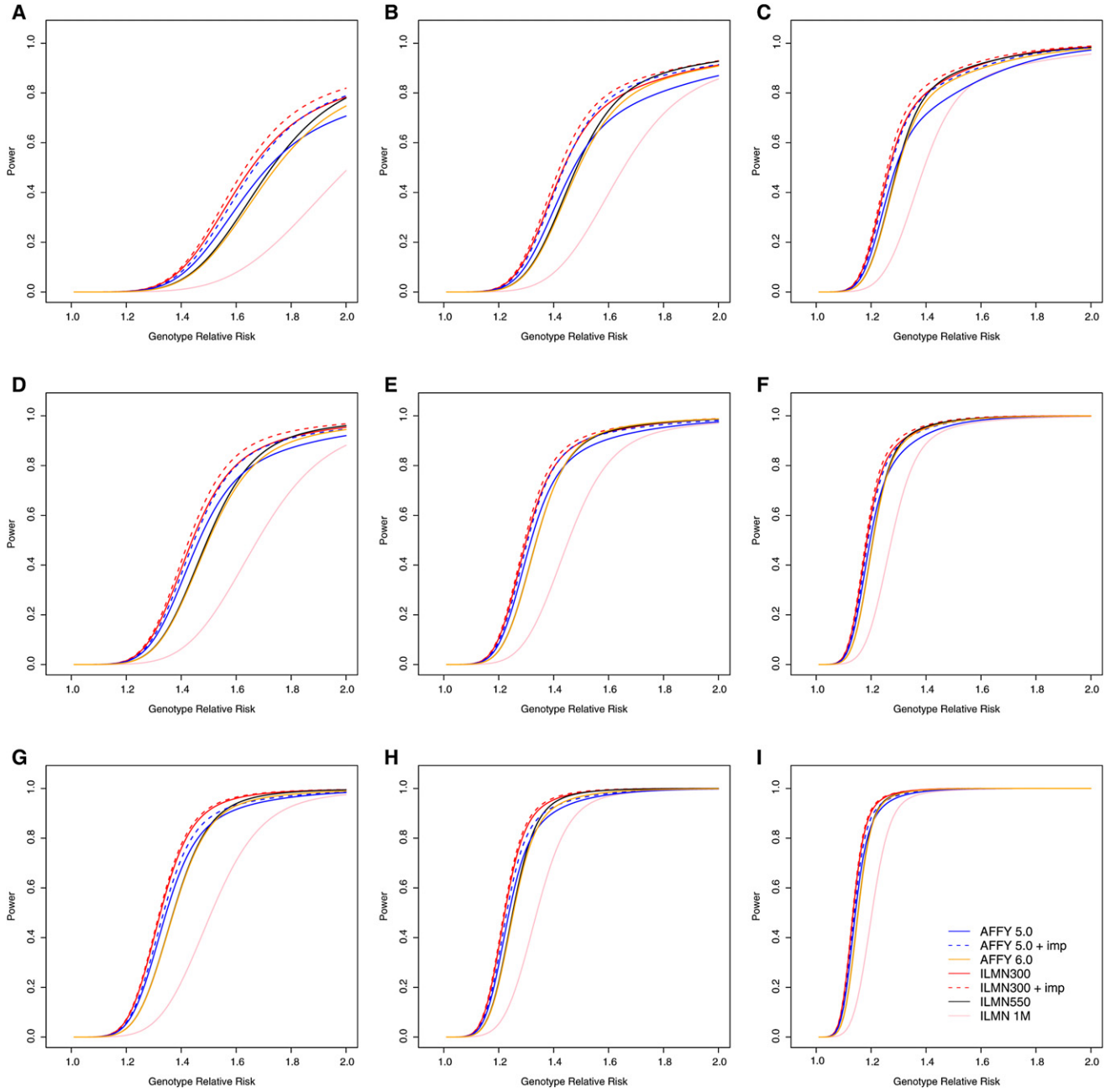

Figure 4. Mean Power to Detect Association to a Disease Where Baseline Sample Size Has Been Varied across Genome-wide SNP Platforms to Reflect Relative Cost

Mean power to detect association ($\alpha = 10^{-5}$) to a disease with a population prevalence of 0.0001 where baseline sample size has been varied across genome-wide SNP platforms to reflect the genotyping cost per sample (sample-size ratios: SNP Array 5.0 = 1.22; HumanHap 300 = 1.32; HumanHap 550 = 1; SNP Array 6.0 = 0.99; HumanHap 1M = 0.57).

RAF ranges are as follows: (A–C) $0.05 \leq \text{RAF} < 0.10$; (D–F) $0.10 \leq \text{RAF} < 0.20$; (G and H) $0.20 \leq \text{RAF} \leq 0.50$. Cases and controls are as follows: (A, D, and G) 1,000 cases, 1,000 controls; (B, E, and F) 2,000 cases, 2,000 controls; (C, F, and I) 5000 cases, 5,000 controls. Mean power was calculated after 10,000 simulations where sample size per simulation for each SNP set was weighted by the maximum r^2 between a randomly selected HapMap SNP (satisfying RAF constraints) and the SNPs on the given genotyping platform (with HapMap release 21 CEU data).

are among populations. However, imputation quality is directly related to how well the haplotype map on which the imputations were based documents variation within the study population. If there is a good haplotype map for the given population, we show that much power can be gained through imputation. However, when a suitable haplotype map is unavailable, it is unlikely that imputations will be of sufficient quality to increase power significantly. Multidimensional scaling can be carried out to compare individuals from a given study to individuals

used to build a haplotype map, thus providing some indication of the suitability of that haplotype map for genotype imputation.

It seems likely that the majority of genotypes that can be accurately imputed for a given data set are previously well covered by the direct genotypes. Therefore, coverage is only increased if the imputed SNPs capture variation at SNPs not previously covered by the SNPs used to carry out the imputation. The largest gains in coverage and power stand to be made at those SNPs that are poorly covered by the direct genotyping, although because of this fact they are significantly less likely to be imputed with sufficient accuracy to pass QC. However, even a small increase in maximum r^2 between an untyped variant and a tagSNP can increase power to associate that particular variant with disease.

A study evaluating the cost efficiency of GWA chips, which did not address imputation, used the sample-size ratio required to equate power as a means of comparing GWA platforms.[13] The authors state when n = 3,000, the HumanHap 300 requires 1.34 times the number of individuals as the HumanHap 550 for power to be equated. However, given that under this scenario the power of the two platforms is 0.929 and 0.957 respectively, there seems to be little need to genotype extra individuals with the HumanHap 300. When power is high, power curves are reasonably flat and large sample-size increases are required for small power gains. Therefore the sample-size ratio required to equate power between platforms is a poor metric on which to choose genotyping platforms because the ratio required to gain practically *equivalent* power is often far less.

Price ratios used here are intended as a guide only. To enable power simulations to be carried out with study-specific sample sizes and price ratios, we have made the R code used to carry out the power simulations available online (see Web Resources). This allows a thorough assessment of which GWA platform is the most cost efficient for a given study. With producers of genome-wide SNP platforms continually offering larger panels at increasing cost, it is a difficult task to accurately assess which platform offers the most value for money. Imputation can greatly reduce interplatform differences in coverage and power, and given current pricing structures, can provide (in combination with direct genotype data) the most value for money in terms of genomic coverage and power to detect association.

Appendix 1. Power Simulations

Power estimations for each SNP set were carried out under the assumption of a multiplicative disease model. For a given baseline sample size on the Illumina HumanHap 550 platform (550), the corresponding baseline sample size for an alternative platform is given by $n_{550} \times CR$, where CR is the cost ratio between the two platforms.

The genotype frequencies of the cases (affected by the disease) and controls (unaffected by the disease) were simulated with the assumption of a disease population prevalence of 0.0001. The functional polymorphism was randomly ascertained by selection of a random HapMap SNP, i, from within a given RAF range ($0.05 \leq \text{RAF} < 0.10$, $0.10 \leq \text{RAF} < 0.20$, $0.20 \leq \text{RAF} \leq 0.50$).We assessed power to detect association to this SNP at $\alpha = 10^{-5}$ by using the above sample sizes and the Cochran-Armitage test for trend. Effective sample size, E_n, which accounts for how well SNP i is covered by the given genome-wide SNP chip, was calculated for each platform and is given by

$$E_n = n \times \max r_i^2,$$

where n is the baseline sample size for the given platform and $\max r_i^2$ gives the maximum r^2 between HapMap SNP i and those on the given SNP set.

The genotype frequencies of the case and control cohorts are as follows:

i	0	1	2
Case	p_0	p_1	p_2
Control	q_0	q_1	q_2,

where i denotes the number of high risk alleles. Genotype frequencies at the SNP in unaffected individuals given the RAF q are

$$q_0 = (1-q)^2,$$
$$q_1 = 2q(1-q), \text{and}$$
$$q_2 = q^2.$$

The disease probability (i.e., the probability that an individual is affected given his or her genotype) is given by

$$\varphi = \frac{k}{\lambda^2 q_2 + \lambda q_1 + q_0},$$

where k is the disease population prevalence, λ is the heterozygous genotype relative risk, and$\lambda^2$is hence the homozygous genotype relative risk. Subsequently, the genotype frequencies in affected individuals are

$$p_0 = \frac{\varphi(1-q)^2}{k},$$
$$p_1 = \frac{2\varphi\lambda q(1-q)}{k}, \text{and}$$
$$p_2 = \frac{\varphi\lambda^2 q^2}{k}.$$

Power is given by the χ^2 distribution with noncentrality parameter (ρ) of

$$\frac{\beta_1 (E_n/2)^2}{\beta_2 - \left(\frac{\beta_3}{E_n}\right)}$$

where
$$\beta_1 = ((q_1 - p_1) + 2(q_2 - p_2))^2,$$
$$\beta_2 = \tfrac{E_n(p_1+q_1)}{2} + 2E_n(p_2+q_2),$$
$$\beta_3 = \left(\tfrac{E_n(p_1+q_1)}{2} + 2E_n(p_2+q_2)\right)^2$$

Acknowledgments

We would like to thank the participants and investigators of the International HapMap and WTCCC projects for generating the

data used herein and for making it freely available to the scientific community. We would also like to thank Cecilia Lindgren for reading and commenting on earlier drafts of the manuscript.

Received: April 16, 2008
Revised: May 30, 2008
Accepted: June 5, 2008
Published online: June 26, 2008

Web Resources

The URLs for data presented herein are as follows:

HapMap, http://www.hapmap.org/
WTCCC, http://www.wtccc.org.uk/
C.A. project homepage, http://www.well.ox.ac.uk/~carl/gwa/cost-efficiency

References

1. Smyth, D.J., Cooper, J.D., Bailey, R., Field, S., Burren, O., Smink, L.J., Guja, C., Ionescu-Tirgoviste, C., Widmer, B., Dunger, D.B., et al. (2006). A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat. Genet. *38*, 617–619.
2. McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D.R., Hinds, D.A., Pennacchio, L.A., Tybjaerg-Hansen, A., Folsom, A.R., et al. (2007). A common allele on chromosome 9 associated with coronary heart disease. Science *316*, 1488–1491.
3. Rioux, J.D., Xavier, R.J., Taylor, K.D., Silverberg, M.S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M.M., Datta, L.W., et al. (2007). Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat. Genet. *39*, 596–604.
4. The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common disease and 3,000 shared controls. Nature *447*, 661–678.
5. Barrett, J.C., and Cardon, L.R. (2006). Evaluating coverage of genome-wide association studies. Nat. Genet. *38*, 659–662.
6. Power, C., and Elliott, J. (2006). Cohort profile: 1958 British birth cohort (National Child Development Study). Int. J. Epidemiol. *35*, 34–41.
7. The International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.
8. The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–862.
9. 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.
10. Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems *58*, 109–130.
11. R Development Core Team (2005). R: A Language and Environment for Statistical Computing (Vienna: Austria).
12. Scherer, S.W., Lee, C., Birney, E., Altshuler, D.M., Eichler, E.E., Carter, N.P., Hurles, M.E., and Feuk, L. (2007). Challenges and standards in integrating surveys of structural variation. Nat. Genet. *39*, S7–S15.
13. Li, C., Li, M., Long, J.-R., Cai, Q., and Zheng, W. (2008). Evaluating cost efficiency of SNP chips in Genome-wide assoiciation studies. Genet. Epidemiol., in press. Published online February 12, 2008. 10.1002/gepi.20312.