Evaluation of genome-wide power of genetic association studies based on empirical data from the HapMap project

Yasuhito Nannya1,2,4, Kenjiro Taura3, Mineo Kurokawa1, Shigeru Chiba2 and Seishi Ogawa2,4,*

1Department of Hematology/Oncology, 2Department of Cell Therapy and Transplantation Medicine, Graduate School of Medicine and 3Department of Information and Communication Engineering, Graduate School of Information Science, University of Tokyo, Tokyo 113-8655, Japan and 4Core Research for Evolutional Science and Technology, Japan Science and Technology Agency, Saitama 332-0012, Japan

Received April 7, 2007; Revised and Accepted July 22, 2007

With recent advances in high-throughput single nucleotide polymorphism (SNP) typing technologies, genome-wide association studies have become a realistic approach to identify the causative genes that are responsible for common diseases of complex genetic traits. In this strategy, a trade-off between the increased genome coverage and a chance of finding SNPs incidentally showing a large statistics becomes serious due to extreme multiple-hypothesis testing. We investigated the extent to which this trade-off limits the genome-wide power with this approach by simulating a large number of case-control panels based on the empirical data from the HapMap Project. In our simulations, statistical costs of multiple hypothesis testing were evaluated by empirically calculating distributions of the maximum value of the $\chi^2$ statistics for a series of marker sets having increasing numbers of SNPs, which were used to determine a genome-wide threshold in the following power simulations. With a practical study size, the cost of multiple testing largely offsets the potential benefits from increased genome coverage given modest genetic effects and/or low frequencies of causal alleles. In most realistic scenarios, increasing genome coverage becomes less influential on the power, while sample size is the predominant determinant of the feasibility of genome-wide association tests. Increasing genome coverage without corresponding increase in sample size will only consume resources without little gain in power. For common causal alleles with relatively large effect sizes [genotype relative risk $\geq 1.7$], we can expect satisfactory power with currently available large-scale genotyping platforms using realistic sample size (~1000 per arm).

INTRODUCTION

Genome-wide association studies have been proposed as a strategy to identify genetic factors with small to moderate genetic effects in the development of human diseases, as typically assumed for a common disease common variant (CDCV) model (1). In this strategy, a disease-associated locus is identified through single nucleotide polymorphisms (SNPs) that show ‘significantly’ different allele frequencies between affected (cases) and unaffected (controls) individuals, and a large number of SNPs are tested for association in an attempt to realistically identify such SNPs (2,3). Although only a theoretical perspective a decade ago (1), with the unprecedented advance in large-scale genotyping technologies (4–6), it has now become a realistic approach to exploring the genetic basis of human disease (7,8). In addition, recent efforts in the International HapMap Project to understand the genetic diversity among human populations (9,10) have greatly contributed to clarifying the extent to which the number of marker SNPs could be reduced to achieve given genome coverage, or how much genome coverage can be obtained with a given marker SNP set by optimally ‘tagging’ untyped SNPs based on the linkage disequilibrium (LD) observed in the human genome (11–16).

*To whom correspondence should be addressed to: Department of Cell Therapy and Transplantation Medicine, The 21st Century COE Program, Graduate School of Medicine, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8655, Japan. Tel: +81 358008741; Fax: +81 358046261; Email: sogawa-tky@umin.ac.jp

© The Author 2007. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
Meanwhile, the major interest of the most researchers, who plan genetic association studies, would be the practical success rates in such attempts and their efficient study designs, rather than mere genome coverage (17,18), because increase in genome coverage might not be linearly translated into gain in power (19,20). In addition, the more SNPs are genotyped to achieve better genome coverage, the higher hurdle is imposed for a target allele to be detected.

This dilemma, known as the trade-off between increased genome coverage and the consequent inflation of null statistics due to extreme multiple testing, is a unique feature of genetic association studies, and is best described by considering the distributions of test statistics for markers truly associated with a causative allele (‘causal distribution’) and for all other markers (‘null distribution’) (21). Regardless of the properties of the causative SNP and whether one or more tagging strategies are used, the null distribution for a given marker set depends on its genome coverage in the study population. In particular, the null distribution with complete genome coverage is related to the overall diversity of the human genome and should substantially shift to the right (7,8,22). On the other hand, for a given disease model, the size of the test statistic expected for the causative SNPs is limited by the number of samples to be analyzed, once they are directly captured by one or more marker SNPs. After all, the feasibility of genome-wide association studies, or the required sample size to obtain realistic power, is determined by the overall diversity of the human genome, or given restricted study resources, the diversity of the human genome determines the property of disease-associated SNPs that can be detected with this approach.

Our questions are, therefore, how diverse is the human genome in view of conducting genome-wide association studies, how much power could be obtained to identify causative SNPs given that diversity and how the typical study parameters affect the power in that situation? To answer these questions, we need to evaluate both null and causal distributions in a quantitative manner. Because both distributions intrinsically depend on the LD structure within N (typically \( \sim 10^5 \)–10) interrelated marker SNPs and the particular location of causative SNPs within the genome, they cannot be calculated in an algebraic manner, but need to be estimated based on the observed data of human genome variations (10,21). So we approach these issues by extensively simulating a large number of case-control panels under both null and alternative scenarios based on the data from the International HapMap Consortiums (9,10), and assess the feasibility and efficient designs of whole genome association studies by estimating the genome-wide power that would be obtained using this genetic approach under varying study conditions.

RESULTS

Estimation of null distributions of the maximum \( \chi^2 \) statistics

In considering the issue of multiple testing in genetic association studies, it is convenient to evaluate the maximum value of the \( \chi^2 \) statistic [max(\( \chi^2 \))] in all the marker SNPs that are truly unrelated to the causative SNP (21). Different statistics can be used (23–26), but the power calculated for this statistic, i.e. the probability of max(\( \chi^2 \)) indicating a true association, will provide a reasonable bottom line to discuss the feasibility of typical genetic association studies (21). When all N marker SNPs are independent, the null distribution for max(\( \chi^2 \)) is given as

\[
\varphi_N(\chi^2) = \frac{d}{\chi^2} \phi(\chi^2)^N,
\]

where \( \phi(\chi^2) \) is the cumulative density function of the \( \chi^2 \) distribution (d.f. = 1). However, since SNPs in real marker sets are variably degenerated due to the presence of LD between adjacent SNPs, we empirically estimated the distribution of max(\( \chi^2 \)) for a series of marker sets by simulating 10 000 null case-control panels, where each panel was generated by randomly resampling phased chromosomes from the HapMap data sets, and max(\( \chi^2 \)) was calculated for each case-control panel. Although the number of resampled chromosomes for each case-control panel (i.e. the sample size) does not significantly affect the distributions (data not shown), there arises some concern about the possibility of underestimating the null distributions due to resampling from very limited numbers of chromosomes, because the latter procedure could restrict the freedom of allelic segregation within the same chromosome. To address this issue, we progressively divided the whole genome into larger numbers of sub-blocks consisting of 10 000 to 10 SNPs in the HapMap Phase II set, and resampled these sub-blocks to simulate distributions of max(\( \chi^2 \)). Reducing the mean block size down to 7.1 kb, these divisions allow for greater freedom of allelic segregation, but does not significantly affect the max(\( \chi^2 \)) distributions until the resampled block size becomes smaller than the mean LD length (27), indicating that our simulations are not likely to substantially underestimate the null distributions (Supplementary Material, Figure S1).

Figure S1 A shows the simulated null distributions in the CEU panel for varying numbers of randomly selected SNPs (‘correlated’ SNP sets). The number of segregating or polymorphic markers contained in each random set is designated as Ns. The theoretical distribution for the same numbers (Ns) of ‘independent’ SNPs, \( \varphi_N(\chi^2) \), is also provided (Fig. 1B). The null distribution increases as the number of randomly selected SNP markers increases, and in a random 1000K set containing 681K segregating SNPs, the threshold \( \chi^2 \) value that provides a genome-wide \( P \)-value of 0.05 or 0.01 becomes as large as 27.6 or 30.5, respectively. On the other hand, reflecting the growing inter-marker LD intensity, the empirical distributions gradually deviate from the theoretical ones, \( \varphi_N(\chi^2) \)’s, for increasing Ns within the corresponding marker sets, underscoring the importance of considering inter-marker LD to avoid overestimation of the statistical threshold for multiple testing, especially for higher marker density.

Evaluation of the inter-marker LD

The intensity of the inter-marker LD in a given marker set is more simply evaluated by fitting the simulated distribution to a theoretical one for independent Nc markers, \( \varphi_N(\chi^2) \) (see Methods). Irrespective of marker sets, fitting is finely
performed except in the vicinity of the maximal points (Supplementary Material, Figure S2). In particular, the distribution in extreme \( x^2 \) values is satisfactorily approximated to provide a rough estimate of the nominal \( P \)-value for given genome-wide thresholds as confirmed by the concordance of the upper \( p \) point in the simulated distribution with the upper \( p/N_c \) point in the \( x^2 \) distribution (d.f. = 1) (Bonferroni) (Table 1). In this formulation, it is reasonable to regard \( N_c \) as the number of hypothetical independent SNPs equivalent to the corresponding marker set, where the null distribution for a large number of mutually degenerated SNPs is described by an integer and the mean intensity of the inter-marker LD is measured through the \( N_c/N_s \) ratio.

\[
n_c \text{ values were calculated for a variety of randomly selected SNP marker sets and plotted against the number of segregating SNP markers therein (Fig. 1C). As the Phase II data contain most of the SNPs in commercially available platforms, including Affymetrix® GeneChip® and Illumina® HumanHap® arrays (28–30), \( N_c \) values were also evaluated for these platforms (Supplemental Material, Table S1). Note that the numbers of segregating SNP markers varies among different HapMap panels, even though the same numbers of SNPs are randomly selected for each panel (Supplementary Material, Figure S3). Figure 1C illustrates how the degree of degeneration within marker SNPs increases in different HapMap panels as more marker SNPs are selected.}
\]
For example, 681K segregating SNPs within a random 1000K set in the CEU panel are equivalent to independent 290K SNPs, indicating that in this panel, these SNPs are degenerated 2.3-fold. On the other hand, the degeneration in 1000K set in the CEU panel are equivalent to independent 290K hypothetical independent SNPs, in the CEU, YRI, and JPT+CHB panels, respectively. Assuming the entire genome shows the similar LD intensity to that in the seven ENCODE regions, and they are equivalent to 1340 (5.8-fold), 2580 (3.8-fold), and 1460 (5.1-fold) hypothetical independent SNPs, in the CEU, YRI, and JPT+CHB panels, respectively. 

The SNPs on the Affymetrix® GeneChip® mapping array sets are degenerated to the same degree as random SNP sets, reflecting the fact that the SNPs on GeneChip® platforms are virtually randomly selected. In contrast, the SNPs on the Illumina® HumanHap300 are selected by efficiently tagging the HapMap Phase I SNPs in CEU, in which redundant SNPs are effectively eliminated (28). As a result, degeneration in the HumanHap300 is substantially reduced compared to the corresponding random marker sets. In CEU, Nc for this 305.1K segregating SNP set (215K Nc) exceeds that for 417.8K segregating SNPs on GeneChip® 500K set (196K), as predicted by the higher genome coverage of the former set (see Table 1 and Supplementary Material, Figure S4). The tagging for CEU also increases the Nc in JPT+CHB, as predicted by the higher genome coverage of the former set (see Table 1 and Supplementary Material, Figure S4). 

<table>
<thead>
<tr>
<th>Platform</th>
<th>Ns</th>
<th>Nc</th>
<th>Fold degeneration</th>
<th>$P = 0.05$</th>
<th>$P = 0.01$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Actual$^{b}$</td>
<td>Actual$^{b}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bonferroni$^{b}$</td>
<td>Bonferroni$^{b}$</td>
</tr>
<tr>
<td>Random 10K</td>
<td>6.8K</td>
<td>6K</td>
<td>1.1</td>
<td>$7.99 \times 10^{-6}$</td>
<td>$1.57 \times 10^{-6}$</td>
</tr>
<tr>
<td>Random 30K</td>
<td>20.6K</td>
<td>17K</td>
<td>1.2</td>
<td>$2.86 \times 10^{-6}$</td>
<td>$5.73 \times 10^{-7}$</td>
</tr>
<tr>
<td>Random 50K</td>
<td>33.9K</td>
<td>27K</td>
<td>1.3</td>
<td>$1.76 \times 10^{-6}$</td>
<td>$4.01 \times 10^{-7}$</td>
</tr>
<tr>
<td>Random 125K</td>
<td>85.1K</td>
<td>60K</td>
<td>1.4</td>
<td>$7.39 \times 10^{-6}$</td>
<td>$1.56 \times 10^{-6}$</td>
</tr>
<tr>
<td>Random 250K</td>
<td>170.7K</td>
<td>105K</td>
<td>1.6</td>
<td>$4.52 \times 10^{-7}$</td>
<td>$9.04 \times 10^{-8}$</td>
</tr>
<tr>
<td>Random 500K</td>
<td>340.4K</td>
<td>179K</td>
<td>1.9</td>
<td>$2.45 \times 10^{-7}$</td>
<td>$5.39 \times 10^{-8}$</td>
</tr>
<tr>
<td>Random 1000K</td>
<td>680.7K</td>
<td>290K</td>
<td>2.3</td>
<td>$1.48 \times 10^{-7}$</td>
<td>$3.41 \times 10^{-8}$</td>
</tr>
<tr>
<td>GeneChip 500K</td>
<td>417.8K</td>
<td>196K</td>
<td>2.1</td>
<td>$2.05 \times 10^{-7}$</td>
<td>$4.94 \times 10^{-8}$</td>
</tr>
<tr>
<td>GeneChip Nsp250K</td>
<td>219.4K</td>
<td>120K</td>
<td>1.8</td>
<td>$3.69 \times 10^{-7}$</td>
<td>$7.94 \times 10^{-8}$</td>
</tr>
<tr>
<td>GeneChip 100K</td>
<td>101.3K</td>
<td>62K</td>
<td>1.6</td>
<td>$7.75 \times 10^{-7}$</td>
<td>$1.38 \times 10^{-7}$</td>
</tr>
<tr>
<td>HumanHap 300</td>
<td>305.1K</td>
<td>215K</td>
<td>1.4</td>
<td>$2.18 \times 10^{-7}$</td>
<td>$4.06 \times 10^{-8}$</td>
</tr>
<tr>
<td>HumanHap 550</td>
<td>513.8K</td>
<td>318K</td>
<td>1.6</td>
<td>$1.41 \times 10^{-7}$</td>
<td>$2.90 \times 10^{-8}$</td>
</tr>
<tr>
<td>HapMap Phase II</td>
<td>2557.4K</td>
<td>603K</td>
<td>4.2</td>
<td>$7.09 \times 10^{-8}$</td>
<td>$1.48 \times 10^{-8}$</td>
</tr>
<tr>
<td>ENCODE 7 regions</td>
<td>7.7K</td>
<td>1.3K</td>
<td>5.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^a$Nominal $P$-value to reach given experiment-wide significance obtained from actual distribution.

$^b$The upper $P$ point of the actual null distribution.

$^c$The argument of the $\chi^2$ distribution (d.f.=1) for cumulative density $1 - P/Nc$.

Causal distribution of max($\chi^2$)

In view of power estimation, our next interest was the expected size of causal distributions relative to that of the inflated null distributions under varying disease/study parameters that affect the former distributions. To illustrate this, we simulated causal distributions of max($\chi^2$) for representative CEU alleles assumed to be causative (Fig. 2). Two thousand case-control panels were generated for each simulation, in which phased HapMap SNPs within 500 Kb around the causative locus were randomly resampled assuming a multiplicative model with varying genotype relative risks (GRRs) and the max($\chi^2$) was calculated for the resampled marker SNPs on GeneChip® 500K. Prevalence of the trait was set to 0.05. While the $\chi^2$ threshold for genome-wide $p$ of 0.05 could inflate from 19.9 for the random 10K set (6K Nc; semivariogram) to as high as 29.8 for complete genome coverage (1023K Nc; dotted lines), these costs of multiple testing are acceptable when LD capture of the causative SNP by one or more markers with high correlation coefficient ($r^2$) can create large causal distributions with practical sample sizes (Fig. 2D–F), i.e. when the causal allele is common.
is trivial (Fig. 2). On the other hand, when causative SNPs with weak to moderate genetic effects are detected with insufficient sample numbers, causal distributions cannot exceed large thresholds resulting from extreme multiple testing, even though more and more SNPs are captured by strong LD. With increasing effect size and sample number, the genome coverage is less influential except for smaller numbers of marker SNPs (<250K). The power gain obtained with increased genome-coverage tends to be offset by the increased cost of multiple testing. After all, in most scenarios, genome coverage is less influential on power when ≥250K random markers or equivalent tag SNPs are used. In contrast, the effect of sample numbers is predominant. To detect weak genetic effects (GRR ≒1.3), the number of samples becomes critical. More than 4000 samples per arm will be required, but the requirement of genome coverage is not substantially increased when more than 250K randomly selected SNPs or their equivalents are used (Fig. 3A). Given a higher genetic effect, this dependence on sample size is dramatically ameliorated, but the genome coverage remains less influential.

### Table 2. The number of corresponding independent markers

<table>
<thead>
<tr>
<th></th>
<th>ENCODEa</th>
<th>Whole genomeb</th>
<th>All Phase IIc</th>
</tr>
</thead>
<tbody>
<tr>
<td>YRI</td>
<td>2580</td>
<td>1971K</td>
<td>1049K</td>
</tr>
<tr>
<td>CEU</td>
<td>1340</td>
<td>1023K</td>
<td>603K</td>
</tr>
<tr>
<td>JPT + CHB</td>
<td>1460</td>
<td>1115K</td>
<td>632K</td>
</tr>
</tbody>
</table>

aNc values calculated for combined SNPs from seven regions.

bNc of ENCODE regions are extrapolated to the entire genome.

cNc of all SNPs in the HapMap Phase II.

(MAF > 0.2) and has a large GRR (>1.7) (Fig. 2A, D and G). In contrast, in the case where the causal allele with smaller MAF value (<0.2) or with a modest to weak GRR (<1.5) is to be detected, the trade-off between increased chance to capture the allele with higher r² using more markers and the accompanying cost of multiple testing can offset the power to varying degrees (Fig. 2A–C, G–I). The effect of ‘collaborative’ capture, i.e. the probability of detecting an association by one of the multiple surrounding marker SNPs other than the SNPs showing max(r²), creates measurable gain in causal distributions and overall power, but does not essentially influence the above observations (Supplementary Material, Figure S5).

### Estimation of genome-wide power

Based on the above consideration, we estimated the genome-wide power in genetic association studies for common (MAF ≥ 0.05) causal alleles with weak to moderate genetic effects. To do this, after assuming all the common SNPs in the human genome being equally causative, we used two sets of SNPs, the Ref ENCODE and the Ref Phase II 5Kb sets (see Methods), as references that are considered as random sampling from the entire SNPs. For each putative causative SNP, we simulated case-control panels as described in the previous section, and calculated the single point power as the proportion of simulated panels whose max(χ²) exceeded a predetermined χ² threshold corresponding to a genome-wide P = 0.01 or 0.05 for each marker set. For genome-wide power, each single point power was averaged for all common SNPs within the reference set. For the Ref Phase II 5Kb set, over-representation of the direct association was adjusted based on the estimated genome coverage of the Phase II data set (see Methods). Figure 3 shows the genome-wide power in the CEU panel that was calculated for the Ref Phase II 5Kb for moderate to small effect sizes (i.e. GRR ≒1.7) assuming various parameter values. The calculation on the Ref ENCODE set provides a largely equivalent estimation of the power (Supplementary Material, Figure S6), although the power is expected to be less reliable for smaller marker sets, reflecting their poor representation of the genome.

Under strong genetic effects (GRR ≥ 2.0) and large sample sizes (≥1500/arm), the power tends to saturate as the number of randomly selected SNPs increases (≥250K), because most of the common SNPs would be already captured by one or more marker SNPs with enough r² (Supplementary Material, Figure S4), and the capture causes large shifts of causal distributions to the extent that the cost of multiple testing

### Power in different HapMap panels and in commercially available platforms

Power is significantly reduced in YRI compared to CEU and JPT+CHB for any marker set (Fig. 4A–C). The lower power in YRI is mainly due to the lower ‘relative’ genome coverage of the marker set (Nc/NcG), rather than the higher cost of type I errors in this population.

The Illumina® HumanHap® series are commercially available platforms that incorporate the tagging theory, in which marker SNPs were selected to efficiently tag the CEU SNPs in the Phase I data set. Tagging seems to be effective, since HumanHap300® in the Ref Phase II 5Kb set shows slightly higher power than the GeneChip® 500K in CEU, although the power is slightly biased by the higher representation of the Phase I SNPs in the Ref Phase II 5Kb set (Fig. 4D). Human-Hap300® shows comparable power to that of GeneChip® 500K, but the power of HumanHap300® is significantly reduced in YRI. In HumanHap550®, more tag SNPs from YRI and JPT+CHB were added to HumanHap300®, the power is more improved in YRI and in JPT+CHB, but the power is also increased to a lesser degree in CEU reflecting a transferability of tag SNPs between CEU and JPT+CHB. The power of various commercially available platforms with various sample sizes are shown in Figure 4E (adaptive threshold) and in Supplementary Material, Figure S7 (fixed threshold). Genome coverage and power of HumanHap550® in the CEU are comparable to those of the random 1000K set (Supplementary Material, Figure S4), an equivalent to Human SNP Array 6.0® that is planned by Affymetrix® (Fig. 4E). Nevertheless, and in spite of the significant difference in cost, the gain of power in HumanHap550® is not so prominent. Also note that the power calculation for Human-Hap550® could be slightly biased by using the subset of the Phase II SNPs as a reference.
Power depends on allele frequencies of causative alleles

Power strongly depends on MAF of causative alleles, and detecting rare causative alleles is very difficult (Fig. 2) (8,20) for two reasons. First, rare variants are difficult to capture in high $r^2$ values. With currently available platforms (GeneChip® 500K or HumanHap550®), most SNPs with more than 0.10 MAF values are captured in high $r^2$, which could be effectively detected in high power given moderate GRRs ($\geq 1.5$) and sample size ($\geq 1000/arm$) (Fig. 5). In contrast, capturing rare causal SNPs (MAF < 0.10) requires many more marker SNPs or their combinations than capturing common SNPs at the more cost of multiple hypothesis testing. Second, even when captured in high $r^2$ with one or more marker SNPs, associations with these rare SNPs are more difficult to detect than those with common SNPs (Fig.5). In common diseases, the existence of multiple pheno-copy variants would further compromise detection (multiple rare variants) (33,34). Thus, regardless of genome coverage, power is consistently lower for less common SNPs (Fig. 6A and C). To detect rare causative SNPs, we need not only to invest in genotyping large numbers of marker SNPs with

Figure 2. Enhancement of causal distributions by various parameters. Combined effects of LD [in max($r^2$)] and effect size (in GRR) on causal distributions under constant sample size (1000/arm) and MAF value (0.225) (A–C), LD and sample size under constant effect size (GRR = 1.5) and MAF value (0.225) (D–F), and MAF and effect size under constant sample size (1000/arm) and LD [max($r^2$) = 1.0] (G–I), are illustrated based on the simulations for six representative CEU alleles analyzed on GeneChip® 500K [rs9782915 in (A and D); rs7543006 in (B and E); rs731030 in (C and F); rs6603803 in (G); rs3052 in (H); rs1307490 in (I)]. Thresholds for genome-wide $P$-value of 0.05 are indicated for random 10K (solid lines), GeneChip 500K (dashed lines), and complete genome coverage (dotted lines), corresponding to $N_c$ values of 6K, 196K, and 1023K ($N_cG$), respectively. Effects of collaborative capture by nearby markers are incorporated, but they are generally small (Supplementary Material, Figure S5).
Figure 3. Genome-wide power of association studies for common causal alleles with weak to moderate genetic effects. Genome-wide power was calculated in CEU by averaging single point power for each putative causal allele over all common (MAF ≥ 0.05) SNPs in the RefPhase II 5K reference set, with increasing marker and sample sizes for small to moderate GRRs (1.3–1.7) in multiplicative disease models. Power was computed using adaptive thresholds for max(χ²) that provides a genome-wide P-value of 0.05 (dark columns) or using a fixed threshold (P = 1 × 10⁻⁵; light columns) for each marker set. The power with an adaptive threshold for a genome-wide P-value of 0.01 was also indicated by a lower bar within each column.
low MAF values by any means, but also to increase the sample size (Fig. 6B and C).

Discussion

Through the current analysis, we empirically determined the size of test statistics for causal as well as null markers under varying degrees of genome coverage and realistic study parameters, and thereby demonstrated how genome-wide power is affected by the interplay between genome-coverage and other determinants. Here it is appropriate to compare the performance [power \((1 - \beta)\) or sensitivity] of the different SNP sets with their specificity (or \(1 - \alpha\)) being constant by applying adaptive thresholds, where \(\alpha\) denotes genome-wide type I error probability. In addition, the power calculated in this way is directly related to false positive report probability (FPRP), which is simply expressed as \(1/[1+(1-\beta)/\alpha]\), which is approximately extended to \(1/[1+m(1-\beta)/\alpha]\) assuming a total of \(m\) independent causative loci having the same effect size. Note that \(\alpha\) is a constant for all SNP sets.
i.e. 0.05 or 0.01. So from our simulations, readers will easily evaluate the power and FPRP expected from given SNP set, sample size and predicted effect size. As long as practical power (for example, $\frac{1}{2}$) is obtained, FPRP is expected to less than 0.5, which will be satisfactory for initial discovery studies.

We estimated genome-wide thresholds based on the simulations using small numbers of HapMap chromosomes. In real studies, the threshold should be determined using their own applicable data sets, where diploid, rather than phased, chromosomes could be used when enough samples are analyzed. A larger number of chromosomes should contain more numbers of rare segregating SNPs, but these rare SNPs would not increase $\chi^2$ thresholds substantially (22).

In terms of the effective number of independent SNPs (Nc) in various marker sets, the diversity of the human genome is likely to be on the order of 1000K in CEU and the corresponding nominal $P$-value giving a genome-wide $\alpha$ error of 0.05 is $5 \times 10^{-8}$. For moderate GRRs (≤1.5), this threshold could be overcome with ≤1500 samples per arm for very common SNPs (MAF > 0.20), but for less common SNPs or those with a small genetic effect (GRR=1.1–1.2), extremely large numbers of samples will be required (Supplementary Material, Figure S8), which urges moves toward sharing typing data across multiple groups as exemplified in recent reports that identified predisposing factors with very modest genetic effects for type 2 diabetes (35–37). The diversity of our genome may not allow for detecting very rare causative alleles (<0.01) with even smaller genetic effects (i.e. GRR < 1.1) using this approach (Fig. 6D).

Under these limitations, several issues should be considered to efficiently exploit study resources and to increase the chance of finding a true association. First, for the increased genome coverage to be effectively translated into power, it needs to be accompanied by a corresponding increase in sample size. When sample numbers are small relative to the effect size, the cost of multiple testing largely offsets the expected increase in the test statistics for causal alleles with
no measurable gain in power, and can even exceed the gain in causal distributions (Fig. 4). Increasing genome coverage with insufficient sample sizes would only consume resources with no substantial benefit in power. In addition, power tends to saturate in higher genome coverage and the effect of increasing the number of marker SNPs is less prominent compared to that of increasing sample sizes. In most simulated situations, more power is expected by doubling the sample size than by doubling the number of maker SNPs. For example, our simulations predict that doubling the sample size using GeneChip® Nsp 250K is almost certainly more efficient than analyzing half of the samples with both Nsp 250K+Sty 250K (Supplementary Material, Figure S9).

The tagging strategy or statistical imputation is effective for increasing genome coverage with limited numbers of marker SNPs (21,38,39), although it does not save the cost of multiple-hypothesis testing. The efficiency of generating a tag SNP set with higher genome coverage, however, is increasingly compromised. The additional gain in power becomes smaller with increasing genome coverage, while more and more effort will be required to find additional independent tag SNPs, because many SNPs are already captured by existing tag SNPs. In addition, we simulated power using ‘All Phase II’ set. In the sense that all references are captured through direct association, this marker set provides the ultimate coverage of the genome. Considering that modest increase of power using ‘All Phase II’ set compared with random 1000K set (Fig. 3), multimarker tagging presumably may not push up the power profoundly. Transferability of a tag SNP set from one population to another is also a problem. Tag SNPs for CEU are transferable to a certain degree to JPT+CEU, but they are less effective for YRI.

In any simulated scenarios, detecting SNPs with lower MAF values (0.05–0.10) is very difficult using whole genome approaches, which is especially true for SNPs with less than 0.05 MAF values. In this situation, genome coverage to capture these rare SNPs becomes definitely important, but the required increase in the sample size is greater for rare SNPs than for common ones. Effort to devising SNP sets for these rare alleles, or exhaustive multimarker tests (21,38), is not likely to be rewarding unless their genetic effects are substantially large.

**MATERIALS AND METHODS**

**HapMap data sets**

The phased genotyping data of the HapMap Phase II (release 21) were obtained from the International HapMap Project web site (http://www.hapmap.org/downloads/phasing/2006-07_PhaseII/) (10). It includes the data from 60 CEU parents (120 chromosomes), 60 YRI parents (120 chromosomes) and the combined set of 45 JPT and 45 CHB unrelated individuals (180 chromosomes), and is provided in three discrete sets (‘all’, ‘consensus’, and ‘phased’), of which we used the former two sets for analysis. The ‘all’ set contains the comprehensive data of all SNPs genotyped in each population including non-segregating sites, and the ‘consensus’ set consists of the intersection of ‘all’ sets from the three population panels. The ‘all’ sets contain 3755 469, 368 5205 and 3776 850 SNPs for CEU, YRI and JPT+CHB, respectively, and the ‘consensus’ set includes 3535 396 SNPs.
Marker sets and the references for power calculation

We generated a series of marker sets consisting of 10K, 30K, 50K, 125K, 250K, 500K and 1000K SNPs, by randomly selecting SNPs from the Phase II ‘all’ sets for each HapMap panel. The number of segregating SNPs in each set is denoted as Ns and shown in Table 1 for CEU panel. Because the Phase II ‘all’ set contains most of the SNPs on commercially available platforms, including Affymetrix® GeneChip® 500K (Nsp+Sty), 250K (Nsp), 100K (Hind+Xba), Illumina® HumanHap300®, and HumanHap550® (Supplementary Material, Table S1), the intersectional SNPs of these platforms with the Phase II ‘all’ set were incorporated into the analysis as representative SNPs of each commercial set. Annotation files for SNPs on GeneChip® series are available from the Affymetrix® web site (http://www.affymetrix.com/products/application/whole_genome.affx). The SNP information of HumanHap® series was kindly provided by Illumina® Inc. A subset of the Phase II SNPs, referred to as ‘RefPhase II 5Kb’, was constructed and used as a reference in the calculation of genome-wide powers by randomly selecting SNPs from the ‘consensus’ set so that each SNP is, on average, 5 Kb apart from the adjacent SNPs. Combined SNPs from the 10 ENCODE regions, denoted as RefENCODE, were used as an alternative reference set. Only common SNPs (MAF ≥ 0.05) were included in the power calculations as putative causal alleles.

Simulation of case-control panels under the null hypothesis

and fitting simulated distributions

Null distributions in genetic association studies are considered for only vaguely defined ensembles having limited population sizes, e.g. all adult Japanese eligible for a study. To obtain asymptotic distributions, we generated 10 000 null case-control panels by randomly resampling phased autosomal chromosomes from the ‘all’ set of CEU, YRI and JPT+CHB. Simulations were performed with different sample numbers, i.e. 500, 750, 1000, 1500, 2000 and 4000 per single arm. For each case-control panel, the maximum $\chi^2$ value (max($\chi^2$), d.f.=1) in the standard allele test was calculated for different marker sets to obtain empirical null distributions of max($\chi^2$).

The simulated distributions, $\Phi(\chi^2)$, were fitted to the null distribution for hypothetical Nc independent SNPs, $\Phi_{Nc}(\chi^2)$, by the least squares method as follows:

$$Nc = \arg \min_{N} \int (\Phi_{N}(\chi^2) - \Phi(\chi^2))^2 d\chi^2$$

The Gnu Scientific Library was used to handle these functions.

Simulation of case-control studies and calculation of power

We consider multiplicative disease models showing a prevalence $e$, and assume a single causative allele whose MAF and GRR are $P$ (≥0.05) and $\gamma$, respectively. Given the penetrance for AA, Aa and aa genotypes as $f_{AA}$, $f_{Aa}$ and $f_{aa}$, respectively, expected genotype frequencies in the case and control panels are given as,

$$P(AA|case) = \frac{p^2 f_{AA}}{e}$$

$$P(Aa|case) = \frac{2p(1-p)f_{Aa}}{e}$$

$$P(aa|case) = \frac{(1-p)^2 f_{aa}}{e}$$

$$P(AA|control) = \frac{p^2 (1-f_{AA})}{1-e}$$

$$P(Aa|control) = \frac{2p(1-p)(1-f_{Aa})}{1-e}$$

$$P(aa|control) = \frac{(1-p)^2 (1-f_{aa})}{1-e}$$

where

$$e = p^2 f_{AA} + 2p(1-p)f_{Aa} + (1-p)^2 f_{aa}$$

$f_{AA} = \gamma^2 f_{aa}; f_{Aa} = \gamma f_{aa}$

According to these allele frequencies, we generated 2000 case-control panels under the alternative hypothesis by resampling a predetermined number of phased chromosomes, and calculated max($\chi^2$) of the marker SNPs for each panel, where the calculations were performed only for those marker SNPs that are within 500 Kb from the putative causal SNP. The proportion of simulated case-control panels whose max($\chi^2$) exceeded the upper 95 or 99% point of the corresponding null distribution for that marker set was defined as the power. The genome-wide power was computed by averaging each power for all SNPs within the reference set. As the number of marker SNPs increases, up to as high as 1000K, there is a considerable chance of detecting direct associations, i.e. the causative SNP is included in the marker set. Assuming 7500K common SNPs within the human genome (17), the Phase II data set includes one-fourth (2167K common SNPs in CEU) of all the common SNPs. Based on this estimation, we excluded three-fourths of the genome-wide powers by randomly selecting SNPs from the ‘all’ set of CEU, YRI and JPT+CHB.

Computational resources

All simulations were run on the GXP clustering computer system in the Department of Information and Communication Engineering, Graduate School of Information Science, University of Tokyo.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.
ACKNOWLEDGEMENTS

This work is totally indebted to the achievement of the International HapMap Consortium and we thank all the people who participated in the project. We also thank Jun Ohashi for helpful discussions.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by Research on Measures for Intractable Diseases, Health and Labor, Sciences Research Grants, Ministry of Health, Labor and Welfare, Research on Health Sciences focusing on Drug Innovation, the Japan Health Sciences Foundation, and Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Agency.

REFERENCES