Published in final edited form as:

Genet Epidemiol. 2008 July ; 32(5): 387–395. doi:10.1002/gepi.20312.

Evaluating cost efficiency of SNP chips in genome-wide association studies

Chun Li^{1,2}, Mingyao Li³, Ji-Rong Long⁴, Qiuyin Cai⁴, and Wei Zheng⁴

¹Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN 37232

²Center for Human Genetics Research, Vanderbilt University School of Medicine, Nashville, TN 37232

³Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104

⁴Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN 37232.

Abstract

Genome-wide association (GWA) studies have recently emerged as a major approach to gene discovery for many complex diseases. Since GWA scans are expensive, cost efficiency is an important factor to consider in study design. However, it often requires extensive and time consuming computer simulations to compare cost efficiency across different SNP chips. Here we propose two simulation-free approaches to cost efficiency comparisons across SNP chips. In the first method, the overall power under a given disease model is calculated for each SNP chip and various sample sizes. Then SNP chips can be compared with respect to the sample sizes required to achieve the same level of power. In the second method, for a desired level of genomic coverage, the effective r^2 threshold values are calculated for each SNP chip. Since r^2 is inversely proportional to the sample size to achieve the same power, the required sample sizes can then be compared among SNP chips. These two methods are complementary to each other. The first approach provides direct power comparisons, but it requires information on disease model and may not be reliable for SNP chips that contain many non-HapMap SNPs. The second approach allows sample size comparisons based on the coverage of SNP chips, and it can be modified for SNP chips that contain non-HapMap SNPs. These methods are particularly relevant for large epidemiological studies in which enough subjects are available for GWA screening and follow-up stages. We illustrate these approaches using five currently available whole genome SNP chips.

Keywords

genome-wide association; cost efficiency; HapMap

INTRODUCTION

Genome-wide association (GWA) studies have now become feasible due to recent developments in genotyping technologies, a rapid decline in genotyping costs [Hirschhorn and Daly, 2005; Wang et al., 2005; Hinds et al., 2005], and the completion of the International HapMap Project [International HapMap Consortium, 2005, 2007]. In GWA studies, investigators typically rely on commercial genotyping products that attempt to provide

adequate coverage across the genome. Whole genome SNP chips from both Affymetrix and Illumina have been successfully used to identify disease genes. Since multiple SNP chips are available, when designing a GWA study, it is necessary to evaluate and compare all available SNP chips and select the one that can provide high quality data in a cost-efficient manner. To date, the comparisons across SNP chips have been mostly on the chips' coverage of the genome [Barrett and Cardon, 2006; Pe'er et al., 2006], irrespective of genotyping cost. Although SNP chips with a higher genomic coverage are more desirable, such chips generally also cost more. Because a GWA scan is expensive, cost efficiency is an important factor to consider in study design [Skol et al., 2006]. However, it has been difficult to compare cost efficiency across different SNP chips without extensive and time consuming computer simulations.

In this paper, we propose two simulation-free approaches to cost efficiency comparisons among SNP chips. Both approaches rely on the information of linkage disequilibrium (LD). For each SNP chip and each SNP in the genome, we calculate LD as measured by r^2 between the SNP and the SNPs on the chip, and obtain the maximum r^2 . The two methods differ in their summarization of the LD information. The first method calculates the power for each SNP and then the overall power for all SNPs across the genome assuming they are equally likely to be associated with the disease. To calculate power for each SNP, information on disease model is needed. This method allows us to compare the cost efficiency of different SNP chips either by comparing power with a fixed sample size or by comparing the required sample sizes with fixed power. In contrast to the first approach, our second approach does not require specification of disease model. Instead, it compares SNP chips through their effective r^2 thresholds at which they achieve the same level of genomic coverage (i.e., fraction of genome covered above an r^2 threshold) [Barrett and Cardon, 2006; Pe'er et al., 2006], or equivalently, the same overall power. In this approach, we compare the cost efficiency of different SNP chips by comparing the required sample sizes when the coverage is fixed. These two methods are complementary to each other. We will describe the methods and illustrate their applications using five currently available SNP chips.

MATERIALS AND METHODS

Relationship between Cost Efficiency and Sample Size Ratio

To calculate the overall genotyping cost for the SNP chips under consideration, we need to know per-subject genotyping cost, which is influenced by multiple cost incurring factors, including the cost of a SNP chip, the time and effort spent on data quality examination, and the time and effort for assay reruns if the result on a sample is declared unreliable, etc. Once these factors are quantified as cost per subject, they can be added to obtain the overall per-subject genotyping cost.

Consider two SNP chips, with per-subject genotyping costs c_1 and c_2 , respectively, with cost ratio $c_1 / c_2 > 1$. Suppose N_1 and N_2 subjects are needed to achieve a certain level of overall power for the two SNP chips, respectively. Obviously, if $N_1 \geq N_2$, the less expensive chip is more cost efficient. We now consider the situation in which $N_1 < N_2$. In this case, the cheaper chip will be more cost efficient when $c_1 N_1 > c_2 N_2$, or equivalently, $N_2 / N_1 < c_1 / c_2$, that is, the sample size ratio N_2 / N_1 is smaller than the cost ratio c_1 / c_2 . Thus, when there is no limitation on the available sample size, the sample size ratio is more relevant to cost efficiency comparisons than the absolute sample sizes. This argument is particularly relevant for large epidemiological studies that have enough subjects for GWA screening and subsequent follow-up stages. For the rest of the paper, we focus on calculation of required sample sizes and sample size ratios for achieving the same power.

Relationship between r^2 and Sample Size

The power of an association test relies on the LD between the test SNP and the disease variant. The LD between two SNPs is often measured by r^2 , which is the square of the correlation coefficient between the two SNPs. If two SNPs are independent of each other, $r^2 = 0$; if they are in perfect LD, $r^2 = 1$. In the latter situation, the SNPs must have the same allele frequencies. If two SNPs have different minor allele frequencies (MAF), then their r^2 must be less than one. Suppose one SNP is the disease variant and the other SNP is a tag SNP. Then, the higher the value of r^2 is, the higher the power is to detect disease association by analyzing the tag SNP. When considering a SNP chip, a SNP in the genome may be tagged by multiple SNPs on the chip at various r^2 levels. For each SNP in the genome, we define its $r_{\max}^2$ as the maximum r^2 at which it is tagged by the SNPs on the chip.

For case-control studies, there is a well-known approximate relationship between sample size and r^2 when the disease model is multiplicative and the risk allele frequencies are similar between the cases and controls, which is expected to hold for most complex diseases [Pritchard and Przeworski, 2001]. This relationship specifies that if N subjects are needed to achieve a certain level of power when the disease variant is directly observed and analyzed, then N/r^2 subjects are needed to achieve the same level of power when analyzing a SNP that tags the disease variant at r^2 . Now consider two SNPs, denoted as SNP1 and SNP2, that tag the disease variant at r_1^2 and r_2^2 , respectively. Then, the sample sizes required for SNP1 and SNP2 to achieve the same power to detect disease association are $N_1 = N / r_1^2$ and $N_2 = N / r_2^2$, respectively. Thus, the required sample size for SNP2 to achieve the same power of SNP1 is $N_2 = N_1 \times r_1^2 / r_2^2$, and the corresponding sample size ratio is $N_2 / N_1 = r_1^2 / r_2^2$. In other words, the sample sizes required to achieve the same power are inversely proportional to the r^2 values.

Comparison of Cost Efficiency through Overall Power of SNP Chips

The first method requires specification of a single-locus disease model. For a SNP chip and given a disease model, we can calculate the power for each known SNP and the overall power averaged over all SNPs in the genome when sample size is fixed. Alternatively, SNP chips can be compared with respect to their sample sizes required to achieve the same level of overall power.

Power for an individual SNP—For a case-control study, if the disease variant is genotyped, power depends on disease allele frequency p , disease prevalence K , genotype relative risks GRR_1 and GRR_2 , sample sizes N_{case} and N_{control} , and significance level α . Given a disease model, the expected allele counts for cases and controls can be calculated and denoted as $n_{0,\text{case}}$, $n_{1,\text{case}}$, $n_{0,\text{control}}$, and $n_{1,\text{control}}$, respectively, where $n_{0,\text{case}} + n_{1,\text{case}} = 2N_{\text{case}}$ and $n_{0,\text{control}} + n_{1,\text{control}} = 2N_{\text{control}}$. Then, the “expected” allele counts under no association can be calculated; for example, the “expected” count of allele 0 in cases under no association is $e_{0,\text{case}} = (n_{0,\text{case}} + n_{0,\text{control}}) \times N_{\text{case}} / (N_{\text{case}} + N_{\text{control}})$. The non-centrality parameter of the allelic chi-squared test is $\lambda = \sum_i (n_i - e_i)^2 / e_i$, where the summation is over the above four categories. The power can then be obtained as $\Pr(\chi^2_{1,\lambda} \geq t)$, where t is the 100(1 - α) percentile of the chi-squared distribution with one degree of freedom. When comparing SNP chips with different numbers of SNPs, different significance levels may be applied to reflect appropriate adjustment of multiple comparisons. Note that instead of genotype relative risks, the disease model can be alternatively specified using odds ratios or population attributable risks.

For a SNP chip, if the disease variant is included on the chip, the power can be calculated as above. When the disease variant is not directly genotyped, then the power will depend on $r_{\max}^2$, the maximum r^2 at which it is tagged by the SNPs on the chip. In this case, $N_{\text{case}} r_{\max}^2$ cases and $N_{\text{control}} r_{\max}^2$ controls would have been needed to achieve the same power if the disease locus had been genotyped.

Overall power—The overall power of a SNP chip can be calculated as the average power over all known SNPs in the genome. In reality, we do not know which SNP is the disease locus. However, if we assume that each SNP is equally likely to be associated with the disease, then we can calculate the overall power, which is the average power over all known SNPs. We note that it would be time consuming and redundant to carry out the power calculation for each SNP. However, since SNPs with the same risk allele frequency and the same $r^2_{\max}$ are expected to have similar power, we can classify all known SNPs into a fine grid of categories defined by MAF and $r^2_{\max}$. For each category c , we can calculate power, f_c , and fraction of SNPs, w_c , falling into this category. Then the overall power is approximately a weighted average $\sum f_c w_c$. This procedure is an extension of the procedure proposed by Jorgenson and Witte [2006]. To compare cost efficiency of different SNP chips, we can compare their overall power when the sample size is fixed. Alternatively, given a disease model, different SNP chips can be compared with respect to the required sample sizes to achieve the same overall power.

Comparison of Cost Efficiency through Effective r^2 Thresholds

In the second method, the effective r^2 threshold values are obtained for a desired level of genomic coverage for the SNP chips under consideration. Since r^2 is inversely proportional to the sample size required to achieve the same power, SNP chips can be compared with respect to their effective r^2 thresholds. This method does not require specification of disease models.

Genomic coverage—The genomic coverage of a SNP chip is often calculated as the fraction of common SNPs (i.e., SNPs with MAF $\geq .05$) in the genome that can be tagged by SNPs on the chip at or above a given r^2 threshold [Barrett and Cardon, 2006; Pe'er et al., 2006]. In other words, it is the fraction of common SNPs with $r^2_{\max}$ no less than the threshold. In general, the higher r^2 threshold, the more stringent the coverage criterion is and the lower fraction of SNPs a chip can cover at that threshold. It is important to evaluate genomic coverage of SNP chips because better coverage is expected to lead to higher power to detect disease variants. A commonly used r^2 threshold value is 0.8.

The genomic coverage can be calculated using the formula of Barrett and Cardon [2006]. Briefly, for a SNP chip and an r^2 threshold, the calculation relies on four values: R (number of common SNPs in the HapMap), T (number of common SNPs on the chip), L (number of common SNPs not on the chip but are tagged at or above the r^2 threshold by a SNP on the chip within ≤ 250 kb), and G (total number of common SNPs in the genome, which is estimated to be 7.5×10^6 [Barrett and Cardon, 2006]). The coverage is estimated as

$$[L / (R - T) \times (G - T) + T] / G. \quad (1)$$

As described by Barrett and Cardon [2006], $L / (R - T)$ computes the fraction of known HapMap common SNPs that are not on the SNP chip but can be tagged by a SNP on the chip. This fraction is multiplied by $G - T$ to estimate the number of common SNPs that are not on the chip but can be tagged by SNPs on the chip. Adding this number by T gives an estimate of the total number of SNPs that are captured by either LD tagging or by inclusion in the chip. Compared to a naive estimate $(L + T) / R$, this formula corrects for over-estimation of the true fraction of captured SNPs and thus is less biased [Barrett and Cardon, 2006]. The formula can also be rewritten as $L / (R - T) + [1 - L / (R - T)] \times T / G$, which indicates that the value of G has little effect on the final estimate as long as T / G , the fraction of common SNPs included in the chip, is small, which is true for current SNP chips and probably for the products in the near future.

Effective r^2 threshold—Alternatively, one can fix the level of desired genomic coverage and calculate the corresponding effective r^2 threshold. The effective r^2 threshold can be

obtained through a grid search. For a fine grid of r^2 threshold values over the interval (0, 1), the genomic coverage for each SNP chip can be calculated. Then, for the desired genomic coverage level, the corresponding effective r^2 threshold can be determined.

Unlike power, genomic coverage does not directly relate to sample size. Thus, no absolute sample sizes are involved in SNP chip comparisons. However, as we explained earlier, sample size ratio is more relevant to cost efficiency comparisons than absolute sample sizes when enough subjects are available. For a desired genomic coverage level, once the effective r^2 threshold values are determined for the SNP chips under consideration, we can then compare the chips by calculating the corresponding required sample size ratios.

We know that when two SNPs tag a disease variant at r_1^2 and r_2^2 , respectively, the sample size ratio required to achieve the same power is $N_2 / N_1 = r_1^2 / r_2^2$. When comparing two SNP chips, we may replace the r^2 values by the effective r^2 thresholds of the chips. That is, the sample size ratio required to achieve the same level of power between two SNP chips may be approximated as $N_2 / N_1 = r_1^2 / r_2^2$, where r_1^2 and r_2^2 are the effective r^2 thresholds for a desired level of genomic coverage of the two chips, respectively. Suppose the per-subject genotyping costs of the two chips are c_1 and c_2 , respectively, with $c_1 > c_2$. Then, as we explained earlier, the second, less expensive chip is expected to be more cost efficient if $N_2 / N_1 < c_1 / c_2$. Using the above approximation, this implies that the second chip is more cost efficient if the per-subject cost ratio is $r_1^2 / r_2^2 < c_1 / c_2$.

Special Handling for SNP Chips with Non-HapMap SNPs

Both methods rely on a large genomic database such as the HapMap. However, each of the two latest SNP chips, Affymetrix SNP Array 6.0 and Illumina Human1M, has about 10% SNPs that are not on the HapMap. According to Affymetrix, the SNP Array 6.0 has 934,968 SNPs, but with 99,854 SNPs (10.7%) not on the HapMap, including 72,379 common SNPs for CEU, 76,016 for CHB, 70,356 for JPT, and 83,412 for YRI. According to Illumina, the Human1M has 1,072,820 SNPs, but with 125,688 SNPs (11.7%) not on the HapMap, including 70,995 common SNPs for CEU, 67,453 for CHB/JPT, and 77,729 for YRI. For these two SNP chips, their power calculation and genomic coverage may be underestimated.

This problem needs to be addressed to allow a fair cost efficiency comparison when one of these two chips is involved. In the first method, power calculation requires information on $r_{\max}^2$ for each SNP, which may be underestimated for these two chips. However, because the magnitude of such underestimation is unknown and probably is highly variable across different SNPs, the bias in the power calculation is difficult to correct.

For the second method, alternative genomic coverage may be estimated for these two SNP chips. We use the SNP Array 6.0 as an example. Suppose there is an “updated HapMap dataset” that consists of the current HapMap SNPs and the SNPs on the SNP Array 6.0. Based on this “updated data”, we can estimate the number of common SNPs, denoted as R_1 , and the number of common SNPs on the chip, denoted as T_1 ; for example, for Caucasians, $R_1 = R + 72,379$ and $T_1 = T + 72,379$. However, due to the lack of LD information between the “new” SNPs and the other HapMap SNPs, we do not know how many HapMap SNPs are tagged by these “new” SNPs, and thus L_1 cannot be directly estimated. However, if we assume that the number of tagged common SNPs that are not on the chip increases proportionally with the number of common SNPs on the chip, that is, $T_1 / T = L_1 / L$, then L_1 can be estimated as $(T_1 / T) \times L$. Therefore, based on the “updated HapMap data”, the alternative genomic coverage of the SNP Array 6.0 can be estimated as

$$[L_1 / (R_1 - T_1) \times (G - T_1) + T_1] / G \quad (2)$$

The original estimate of genomic coverage in (1) ignored the non-HapMap SNPs that are on the SNP Array 6.0 and can be viewed as a “lower bound” of the coverage, whereas the coverage in (2) may provide a more appropriate estimate. The alternative coverage estimate for the Human1M can be similarly calculated.

We note that this alternative coverage calculation is only a temporary solution to the current situation in which a substantial fraction of SNPs on these two chips are not yet included in the HapMap. We will be able to calculate the coverage of these two SNP chip using coverage definition in (1) once LD information between these non-HapMap SNPs and the HapMap SNPs is publicly available.

RESULTS

We illustrate the methods using five currently available SNP chips: Affymetrix SNP Array 5.0 (500,568 SNPs) and SNP Array 6.0 (934,968 SNPs), and Illumina HumanHap300 (317,511 SNPs), HumanHap550 (555,352 SNPs) and Human1M (1,072,820 SNPs). All the calculations were based on the HapMap data (International HapMap Consortium, 2005; www.hapmap.org). Allele frequency and LD data for the four HapMap populations (CEU, CHB+JPT, YRI) were obtained from HapMap release #22.

Comparison of Cost Efficiency through Overall Power of SNP Chips

For each SNP chip, we calculated the overall power for various levels of prevalence ($K = 0.01, 0.05, 0.10$), genotype relative risks (multiplicative effect with $GRR = 1.25, 1.50, 1.75$), and sample size ($N = N_{\text{case}} = N_{\text{control}} = 1,000$ to 10,000 with increment of 10). For each SNP, the minor allele was designated as the risk allele. The significance level was $0.05/M$, where M is the number of SNPs on the chip. Figure 1 displays the overall power of the five SNP chips under multiplicative disease models with prevalence $K = 0.05$. We note that the results for the SNP Array 6.0 and the Human1M are underestimated due to ignorance of non-HapMap SNPs in the power calculation. As expected, the overall power increases as sample size or GRR increases. For a disease model with $GRR = 1.25$, all five chips have low to moderate overall power even for sample sizes as large as $N = 4,000$. The power improves as the genetic effect of the disease variant increases. When $GRR = 1.50$ and $N = 2,000$, the power ranges from 0.80 for the SNP Array 5.0 to >0.89 for the Human1M for studies with Caucasian subjects. The results are similar for the CHB+JPT samples but have a wider range for the YRI sample, from 0.58 for the HumanHap300 to >0.79 for the Human1M. When $GRR = 1.75$, all five SNP chips have >0.80 overall power for the CEU and CHB+JPT samples even with sample sizes as small as $N = 1,000$.

To compare cost efficiency of different SNP chips, we used the HumanHap550 as a reference chip. For each sample size, we calculated the sample sizes for the other four chips required to achieve the same overall power as the HumanHap550, and the corresponding sample size ratios (Figure 2). For the CEU and CHB+JPT samples, when the disease model has $GRR = 1.25$, because the power is similar among the five chips, the sample size ratios are near one when the sample size for the HumanHap550 is $N \leq 4,000$. When $GRR = 1.50$, the differences in overall power among the chips begin to manifest. For example, for studies with Caucasian subjects, the overall power is 0.96 for the HumanHap550 when $N = 3,000$. To achieve the same level of overall power, N must be 5,870 for the SNP Array 5.0 (sample size ratio 1.96) and 4,030 for the HumanHap300 (sample size ratio 1.34). Unfortunately, the results for the other two latest chips are underestimated due to ignorance of non-HapMap SNPs in the power calculation. When the sample size for the HumanHap550 is very large, the overall power is close to one. In this situation, although a larger sample size is still required for the low-power chips to achieve the same overall power, the sample size ratio starts to reduce. This phenomenon can be seen in Figure 2(C), but it exists for any disease model when the sample size is large

enough. For other prevalence levels, although the absolute values of the overall power were different, the results had similar patterns of sample size ratios (data not shown).

Comparison of Cost Efficiency through Effective r^2 Thresholds

For the second method, we calculated the genomic coverage of these five SNP chips at various r^2 threshold values ($r^2 = 0.1$ to 1.0 with increment of 0.001). Figure 3(A) displays the genomic coverage of the five SNP chips at $r^2 = 0.7$ to 0.9 . For illustration purpose, we again used the HumanHap550 as the reference chip. For a range of target coverage levels for the HumanHap550, we obtained the ratios of effective r^2 thresholds required to achieve the same level of genomic coverage between each of the other four chips and the HumanHap550 (Figure 3B). For each of the four chips, its curve represents r_1^2 / r_2^2 at various genomic coverage levels, where r_1^2 and r_2^2 are the corresponding effective r^2 thresholds for the HumanHap550 and the chip, respectively. At a given coverage level, if the per-subject cost ratio c_1 / c_2 between the HumanHap550 and the chip is above the curve, that is, $c_1 / c_2 > r_1^2 / r_2^2$, then the chip is more cost efficient than the HumanHap550; otherwise, the HumanHap550 is more cost efficient.

For example, for the HapMap CEU sample, the HumanHap550 covers 87% of the genome at $r_1^2 = 0.8$, whereas the SNP Array 5.0 achieves the same coverage at $r_2^2 = 0.383$. Thus, for a study using Caucasian samples, $r_1^2 / r_2^2 = 0.8 / 0.383 = 2.09$ times as many subjects are needed for the SNP Array 5.0 to achieve the same level of power as the HumanHap550. Similarly, the sample size ratio for the HumanHap300 is 1.38. These results are similar to those obtained through the power calculation.

The SNP Array 6.0 achieves the same coverage at $r_2^2 = 0.617$ when coverage was calculated using (1) and $r_2^2 = 0.803$ when coverage was calculated using (2). Note that the former is an underestimate because of the ignorance of non-HapMap SNPs on the chip, while the latter may be more appropriate. For a study using Caucasian samples, the result from (2) suggests that the SNP Array 6.0 and the HumanHap550 have very similar genomic coverage and will achieve equal power with the same number of subjects (i.e., $r_1^2 / r_2^2 = 0.8 / 0.803 = 1.00$). In contrast, the result from (1) suggests that $r_1^2 / r_2^2 = 0.8 / 0.617 = 1.30$ times as many Caucasian subjects are needed for the SNP Array 6.0 as that for the HumanHap550. Similarly, for the CHB+JPT samples, the required sample size ratio is 1.09 under (1) and 0.92 under (2). For the YRI sample, the SNP Array 6.0 has a universally higher coverage than the HumanHap550, and the required sample size ratio is 0.91 under (1) and 0.87 under (2).

DISCUSSION

The GWA approach has been successfully applied in studies of type 2 diabetes [Sladek et al., 2007; Saxena et al., 2007; Zeggini et al., 2007; Scott et al., 2007], breast cancer [Stacey et al., 2007; Hunter et al., 2007; Easton et al., 2007], prostate cancer [Gudmundsson et al., 2007; Yeager et al., 2007], and other diseases [Wellcome Trust Case Control Consortium, 2007; Moffatt et al., 2007; Hakonarson et al. 2007; Samani et al., 2007]. It will undoubtedly be a major approach for gene discovery of many complex diseases. Several SNP chips are available using either the Affymetrix or the Illumina platform, with different properties of genomic coverage and a wide range of prices. Despite these differences, both platforms have been successfully employed to identify novel disease genes, and the results are mostly consistent across studies using different platforms. For example, among the four published type 2 diabetes GWA studies, two studies used Affymetrix chips [Saxena et al., 2007; Zeggini et al., 2007] and the other two used Illumina chips [Sladek et al., 2007; Scott et al., 2007].

When designing a GWA study, it is necessary to evaluate and compare all available SNP chips and select the one that is the most cost efficient. Evaluation of cost efficiency allows investigators to minimize the overall genotyping cost while maintaining the same levels of

genomic coverage and power. On the other hand, when the total budget is fixed, using the most cost efficient SNP chip will allow investigators to increase the sample size and thus the power to detect disease variant. However, cost efficiency comparisons have been difficult without time consuming computer simulations. In this paper, we proposed two simulation-free methods. The first method is based on direct power calculation, which is desirable if disease model information is available. However, it underestimates the power for SNP chips with a significant fraction of non-HapMap SNPs, such as the SNP Array 6.0 and the Human1M. The second method is based on the genomic coverage of SNP chips and does not require specification of disease models. Although these two methods compare SNP chips from different perspectives, they are complementary to each other and may provide similar results. For both methods, we used the HumanHap550 as a reference chip for the purpose of illustration. Similar cost efficiency comparisons can be carried out between any two SNP chips, including future, denser chips. Although the two methods perform cost efficiency comparisons from different perspectives, the ranks of SNP chips based on their overall power are consistent with the ranks based on their genomic coverage.

Both methods rely on a reference database to provide LD information between SNPs in the genome and the SNPs on a chip. Currently, only the HapMap database is publicly available for our calculation. It has been reported that the HapMap SNPs may have a major ascertainment bias [Clark et al., 2005], which might restrict their applicability to many biological questions. For example, less frequent variants are under-represented in the HapMap, and this is partly why SNP chip coverage reported in the literature has been focused mostly on common SNPs with $MAF \geq 0.05$. For the same reason, we restricted our calculations to common SNPs as well. As such, our results may apply well to common SNPs and may or may not apply to SNPs with less frequent alleles.

Because the methods rely on a reference database such as the HapMap, problems arise for the Affymetrix SNP Array 6.0 and Illumina Human1M, for which a substantial fraction of SNPs on the chip are not on the current HapMap. Ignoring non-HapMap SNPs will lead to underestimation of the overall power and genomic coverage and an unfair comparison if such a SNP chip is involved. This problem is difficult to address for the first method because little information is available on the magnitude of $r^2_{\max}$ underestimation. For the second method, we proposed an alternative coverage estimate, which takes the non-HapMap SNPs into account and may be more appropriate than the estimate that is solely based on the HapMap SNPs. We note that this alternative coverage calculation should be viewed as a temporary solution to the current situation in which a substantial fraction of SNPs on these two chips are not yet included in the HapMap.

To allow feasible comparisons among genome-wide SNP chips, both methods require some assumptions and approximations. Although the assumptions and approximations are reasonable, exact accuracy may not be achieved, and the results may vary depending on what disease model or what target level of genomic coverage one would choose. Thus, researchers should interpret the results with caution. A prudent approach is to carry out cost efficiency comparisons at various reasonable disease models using the first method and/or various levels of desired genomic coverage using the second method, and determine if the results are consistent across different conditions.

To allow comparisons and joint analyses between data sets using different SNP chips, genotype imputation methods have been developed [Nicolae, 2006; Marchini et al., 2007], which often rely on a reference database such as the HapMap to provide the LD information between genotyped and ungenotyped SNPs. Genotype imputation will certainly help reduce the differences between two SNP chips to some extent. However, the accuracy of the imputed data depends critically on both the coverage of the chip and the accuracy of the LD information

from the reference database. If a SNP chip has low coverage for a region due to weak LD, the reference database will not provide much information and the accuracy and coverage of the imputed data for the region will be low. If another SNP chip has a better coverage for the same region, fewer SNPs need to be imputed and the imputed data will also have a higher level of accuracy. Thus, the quality of imputed data will partly depend on what SNP chip to start with. Moreover, the LD information provided by the current HapMap is estimated based on a sample of 60 to 90 subjects for each HapMap population. The inevitable variation in LD estimation [Terwilliger and Hiekkalinna, 2006] will also have an effect on the accuracy of the imputed data.

The method presented in this paper can be applied not only to the material cost of SNP chips, but also to the cost of other factors that are important when designing GWA studies, such as the time and effort spent on running assays, data quality examination, and assay reruns if necessary. All these factors can be reflected in the calculation of per-subject genotyping cost, which can then be used for cost efficiency comparisons across SNP chips. Our approach is particularly relevant for large epidemiological studies in which enough subjects are available for GWA screening and follow-up stages when budget allows. For rare diseases or studies without large enough sample size, the SNP chip with the highest genomic coverage may be preferred to achieve high power to detect disease association. We recognize that there are other factors that investigators may take into account when selecting SNP chips, including coverage over genes or regions of interest [Li et al., 2008] and the capability of inferring copy number variations (CNVs). Although considerations such as gene coverage and CNV inference can also be viewed as measures of “efficiency”, they are more difficult to quantify.

Our method is computationally fast. It took only a few hours to carry out the grid calculation of power and genomic coverage. Moreover, the results of genomic coverage for the current SNP chips can be stored, allowing for comparisons with any future SNP chips. As more GWA studies are being planned and carried out, we believe the methods are a timely tool for cost efficiency comparisons, which will help investigators design more efficient GWA studies. More results can be obtained from <http://biostat.mc.vanderbilt.edu/SNPChipCostEfficiency>

Acknowledgment

We thank the Editor and an anonymous reviewer for their critical comments, which have greatly improved the manuscript. This research was supported in part by NIH grants RO1CA64277 and RO1CA90899 (to CL, JRL, QC, WZ), and by a University Research Foundation grant and McCabe Pilot Award from the University of Pennsylvania (to ML).

References

- Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. *Nat Genet* 2006;38:659–662. [PubMed: 16715099]
- Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* 2005;15:1496–1502. [PubMed: 16251459]
- Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struwing JP, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 2007;447:1087–1093. [PubMed: 17529967]
- Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* 2007;39:631–637. [PubMed: 17401366]
- Hakonarson H, Grant SFA, Bradfield JP, Marchand L, Kim CE, Glessner JT, Grabs R, et al. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature* 2007;448:591–594. [PubMed: 17632545]

- Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. Whole-genome patterns of common DNA variation in three human populations. *Science* 2005;307:1072–1079. [PubMed: 15718463]
- Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005;6:95–108. [PubMed: 15716906]
- Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, et al. A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* 2007;39:870–874. [PubMed: 17529973]
- The International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;437:1299–1320. [PubMed: 16255080]
- The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449:851–861. [PubMed: 17943122]
- Jorgenson E, Witte JS. Coverage and power in genomewide association studies. *Am J Hum Genet* 2006;78:884–888. [PubMed: 16642443]
- Li M, Li C, Guan W. Evaluation of coverage variation of SNP chips for genome-wide association studies. *Eur J Hum Genet*. 2008(in press)
- Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;39:906–913. [PubMed: 17572673]
- Moffatt MF, Kabisch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, et al. Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature* 2007;448:470–473. [PubMed: 17611496]
- Nicolae DL. Testing untyped alleles (TUNA) - applications to genome-wide association studies. *Genet Epidemiol* 2006;30:718–727. [PubMed: 16986160]
- Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet* 2006;38:663–667. [PubMed: 16715096]
- Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001;69:1–14. [PubMed: 11410837]
- Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, et al. Genomewide association analysis of coronary artery disease. *New Engl J Med* 2007;357:443–453. [PubMed: 17634449]
- Saxena R, Voight BF, Lyssenko V, Burt NP, de Bakker PI, Chen H, Roix JJ, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007;316:1331–1336. [PubMed: 17463246]
- Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007;316:1341–1345. [PubMed: 17463248]
- Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006;38:209–213. [PubMed: 16415888]
- Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007;445:881–885. [PubMed: 17293876]
- Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, Masson G, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* 2007;39:865–869. [PubMed: 17529974]
- Terwilliger JD, Hiekkalinna T. An utter refutation of the "Fundamental Theorem of the HapMap". *Eur J Hum Genet* 2006;14:426–437. [PubMed: 16479260]
- Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005;6:109–118. [PubMed: 15716907]
- The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447:661–678. [PubMed: 17554300]
- Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* 2007;39:645–649. [PubMed: 17401363]

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 2007;316:1336–1341. [PubMed: 17463249]

Figure 1. Overall power for the five SNP chips under disease models with (1A) $GRR = 1.25$, (1B) $GRR = 1.50$, (1C) $GRR = 1.75$. All models have prevalence $K = 0.05$. The results for the SNP Array 6.0 and the Human1M are underestimated due to ignorance of non-HapMap SNPs in the power calculation.

Figure 2.

Comparison of cost efficiency based on overall power between the HumanHap550 and the other four chips. Plotted is the sample size ratio for each of the other four chips required to achieve the same overall power as the HumanHap550 with sample size at the x-axis. All disease models have prevalence $K = 0.05$ and (1A) $GRR = 1.25$, (1B) $GRR = 1.50$, (1C) $GRR = 1.75$. The results for the SNP Array 6.0 and the Human1M are underestimated due to ignorance of non-HapMap SNPs in the power calculation.

Figure 3.

(A) Genomic coverage of the five SNP chips at various r^2 thresholds. The dashed curves are based on our alternative estimation of coverage for the SNP Array 6.0 and the Human1M, accounting for non-HapMap SNPs. (B) Comparison of cost efficiency between the HumanHap550 and the other four chips using genomic coverage. Plotted is the ratio of effective r^2 thresholds, r_1^2 / r_2^2 , where r_1^2 is the r^2 threshold for HumanHap550 for a target coverage level (x-axis) and r_2^2 is the r^2 threshold for the corresponding SNP chip to achieve the same level of coverage. Note that the x-axis is not evenly spaced with respect to coverage levels.