

Efficiency and power in genetic association studies

Paul I W de Bakker^{1-4,8}, Roman Yelensky^{1,2,5,8}, Itsik Pe'er^{1,4}, Stacey B Gabriel⁴, Mark J Daly^{1,4,6} & David Altshuler^{1-4,6,7}

We investigated selection and analysis of tag SNPs for genome-wide association studies by specifically examining the relationship between investment in genotyping and statistical power. Do pairwise or multimarker methods maximize efficiency and power? To what extent is power compromised when tags are selected from an incomplete resource such as HapMap? We addressed these questions using genotype data from the HapMap ENCODE project, association studies simulated under a realistic disease model, and empirical correction for multiple hypothesis testing. We demonstrate a haplotype-based tagging method that uniformly outperforms single-marker tests and methods for prioritization that markedly increase tagging efficiency. Examining all observed haplotypes for association, rather than just those that are proxies for known SNPs, increases power to detect rare causal alleles, at the cost of reduced power to detect common causal alleles. Power is robust to the completeness of the reference panel from which tags are selected. These findings have implications for prioritizing tag SNPs and interpreting association studies.

Complete genome sequencing offers a comprehensive approach to test all human genetic variation for association to clinical traits. Although routine sequencing of thousands of genomes remains impractical, it has become possible to test systematically most human heterozygosity that is due to common genetic variations^{1,2}. Correlations among nearby variants (linkage disequilibrium, LD) can improve the cost-effectiveness of such studies³⁻⁵, guiding selection of informative 'tag' SNPs⁶ and providing information about nearby variants not genotyped. The International HapMap Project is a resource that provides empirical genome-wide data to support such analyses⁷⁻⁹.

Given practical limitations on genotyping, investigators are forced to make several practical decisions: (i) selecting and prioritizing tag SNPs¹⁰⁻¹⁹, (ii) deciding which tests of association to use²⁰⁻²⁶ and (iii) evaluating statistical significance of putative findings²⁷⁻²⁹ (**Box 1**). Genotyping a higher density of tag SNPs increases the fraction of sites captured through LD³⁰, but the quantitative relationship between additional genotyping and increased power in association studies is not well described. The use of multimarker haplotypes shifts this relationship toward greater efficiency³¹ but has certain drawbacks: if haplotype testing increases the degrees of freedom or number of tests in statistical analysis, it may decrease, rather than increase, overall power²⁴. Many studies will rely on data from the International HapMap Project, which is an extensive but incomplete inventory of common genetic variation⁸. Therefore, it is crucial to understand how tags selected from HapMap compare in power to those selected from a more comprehensive resource.

We set out to study the trade-offs between efficiency and power for different tagging and testing approaches. Because expected power in disease association studies is the most relevant measure of merit (e.g., compared with the distribution of correlation coefficients (r^2) between tag SNPs and untyped variants), we explicitly modeled disease association studies. Second, because varying both the density of tag SNPs and the statistical testing procedure can influence the number of statistical tests (and many of these tests are not independent), we empirically assessed significance thresholds. Finally, because results depend intimately on the true properties of human LD, which are not necessarily well-modeled by population-genetic simulations³², we carried out these evaluations using empirical (rather than simulated) human genotype data.

RESULTS

Disease association studies using empirical genotype data

We started by creating case-control panels using empirical genotype data from the HapMap ENCODE project⁸. Ten 500-kb ENCODE regions were resequenced in 48 individuals, and all discovered SNPs (as well as any others in dbSNP) were genotyped in 269 HapMap samples: 30 parent-offspring trios from the Yoruba people in Ibadan, Nigeria (YRI); 30 parent-offspring trios from Utah, USA, with northern and western European ancestry (from the Centre d'Etude du Polymorphisme Humain; CEU); 45 unrelated Han Chinese people from Beijing, China (CHB); and 44 unrelated Japanese people from Tokyo, Japan (JPT). This data set contains 16,970 SNPs (one every

¹Center for Human Genetic Research and ²Department of Molecular Biology, Massachusetts General Hospital, 185 Cambridge Street, CPZN-6818, Boston, Massachusetts 02114-2790, USA. ³Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. ⁴Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. ⁵Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts, USA. ⁶Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA. ⁷Diabetes Unit, Massachusetts General Hospital, 185 Cambridge Street, CPZN-6818, Boston, Massachusetts 02114-2790, USA. ⁸These authors contributed equally to this work. Correspondence should be addressed to M.J.D. (mjdaly@chgr.mgh.harvard.edu) or D.A. (altshuler@molbio.mgh.harvard.edu).

Received 14 July; accepted 27 September; published online 23 October 2005; doi:10.1038/ng1669

BOX 1 TERMINOLOGY

Variants to be tested for association to phenotype are termed **putative causal alleles**.

The hypothetical relationship between alleles and phenotypes is termed the **genetic model**.

Genotype data used to guide experimental design (selection of tag SNPs and definition of statistical tests to be done) is termed the **reference panel**; HapMap is one such panel.

Tags are the subset of variants genotyped in a disease study. SNPs that are not typed in the study but whose effect can be studied through LD with a tag are termed **proxies**. A tag with perfect correlation ($r^2 = 1$) to an untyped putative causal allele is termed a **perfect proxy**.

The allelic hypotheses examined for association to disease (based on genotypes of the tags) are termed **tests**. A test that is simply the allele of a tag is termed a **single-marker test**. Tests based on combinations of tags are termed **multimarker tests**. A **specified multimarker test** examines a particular allelic combination (a haplotype) of multiple tags based on its observed correlation to a putative causal (untyped) allele in the reference panel. An **exhaustive multimarker test** searches over many or all allelic combinations of tags in the hope of finding a test that captures a hitherto unseen putative causal allele.

~300 bp) with an allele frequency distribution that is almost complete for common alleles; it is available from the HapMap project website.

To simulate a case-control panel, we designated one SNP from this data set to be 'causal'. We calculated an effect size such that if this SNP were directly tested in 1,000 cases and 1,000 controls, power would be 95% to achieve a nominal P value of 0.01. Because our concern was the relative effect on power of tagging and analysis strategies (rather than absolute power), and to make it possible to average results over all putative causal alleles, we fixed the absolute power for each putative causal SNP. Constant power requires minor allele frequency to be inversely correlated to penetrance: in this model, rare alleles are assigned a stronger effect than common alleles (**Supplementary Fig. 1** online). This approach avoids consideration of uninformative scenarios where power is uniformly high (such that any tagging strategy might suffice) or nonexistent (such that tagging is irrelevant).

To simulate the case-control studies, we drew chromosomes spanning each 500-kb region at random from the phased empirical data, conditional on the genotype and effect size at the causal SNP. We repeated this step until there were 1,000 cases and 1,000 controls in each panel and then created 25 such panels for each causal SNP. Finally, we repeated the entire process over all SNPs in the data, generating a large collection of case-control panels in which each SNP has an equal chance of being causal.

Figure 1 Distributions of the test statistic in a typical ENCODE region. Maximum χ^2 statistics for association to disease status are evaluated in the simulated case-control panels (solid line) and random null panels (dotted line). The study-wide significance threshold (vertical gray line) is empirically determined such that the maximum χ^2 test statistic exceeds it in 1% of the null panels (region-wide $P = 0.01$). True associations in the simulated case-control panels with a test statistic below the threshold are rejected (false negatives). Owing to the empirical multiple testing correction, absolute power to detect an association drops from 95% (nominal) to 60% (YRI) and 68% (CEU and CHB+JPT), averaged over all ten ENCODE regions.

We selected tag SNPs and defined statistical tests from a reference panel under a variety of scenarios. We evaluated association for each statistical test using standard 2×2 χ^2 comparisons of cases and controls. The significance threshold for declaring association was based on the empirical null distribution (**Supplementary Note** online): the tags and statistical tests selected in each scenario were examined in a set of null panels (in which no SNP is causal), with the maximum χ^2 value exceeded in 1% of null panels chosen as the threshold to declare a positive result (region-wide corrected P value of 0.01). We report the proportion of case-control panels in which an association was detected, averaged over all putative causal SNPs and over all ENCODE regions.

Capturing all sites observed in a complete reference panel

We began by examining the relationship between the number of SNPs genotyped and statistical power in the best-case scenario: where complete resequencing has been done in a reference panel, such that all putative causal alleles have been observed. We first examined only common alleles: we selected tags to capture alleles with frequency $\geq 5\%$ in the reference panel and limited the set of putative causal alleles in the simulations to those with a frequency $\geq 5\%$.

Figure 1 shows the distribution of the maximum χ^2 values in all null panels and in all causal panels. Nominal power is set to 95% if each causal site is examined as a single test, but the average power after testing all common sites in each 500-kb region falls to 60% (YRI) and 68% (CEU and CHB+JPT). This decline simply represents the power loss resulting from an empirical correction for having tested many hundreds of SNPs in each 500-kb region; the decline in power tracks with the extent of LD in each set of DNA samples.

The simplest and most conservative approach to selecting tag SNPs is to select a subset of nonredundant SNPs from the reference panel such that every common allele either is directly genotyped or has a perfect proxy ($r^2 = 1.0$) among the tags. The reduction in the number of genotypes required (compared with testing all common SNPs directly) was 46% (YRI) and 65% (CEU and CHB+JPT; **Fig. 2**). Because all sites are perfectly captured, relative power remained at 100% compared with testing all common causal alleles directly. (Hereafter, we report the 'relative power' of each tagging strategy: power under a given tagging-testing strategy compared with that obtained by testing all common sites directly.)

We next asked whether multimarker (haplotype) tests could improve the genotyping efficiency, as previously proposed^{10,31}. Because we were concerned about loss of power due to the introduction of additional statistical tests, we developed a strategy in which an identical set of tests of association with one degree of freedom (d.f.) are done but we allow a haplotype of tags to serve as surrogate for an untyped SNP (rather than restricting statistical tests to genotypes of single tags). In other words, if a specific multimarker combination (*i.e.*, haplotype of tag SNPs) can serve as an effective proxy for some

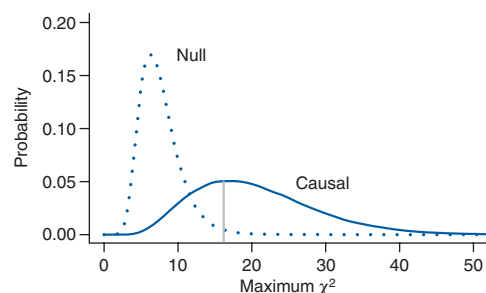
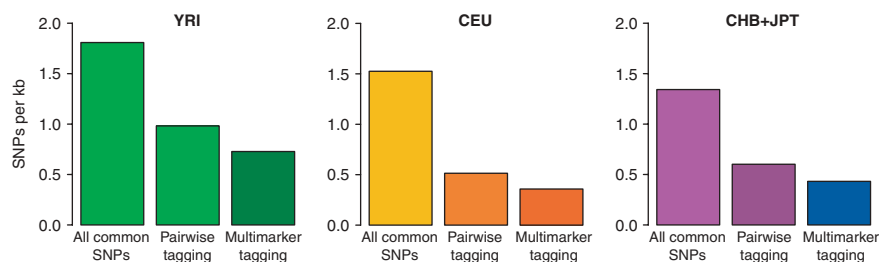


Figure 2 Efficiency afforded by a tagging approach. Using a pairwise tagging and single-marker analysis strategy, a nonredundant ($r^2 = 1$) subset of all SNPs provides 100% relative power to capture all common SNPs ($\geq 5\%$) in the ENCODE data. Efficiency is increased further, while retaining 100% relative power, by the use of multimarker haplotypes.



putative causal alleles, then these alleles need not be typed as tag SNPs (or tested as single markers). In this method, each single tag, as well as each specific haplotype defined above, is tested for association. To avoid overfitting, we required the tags in a specified multimarker test themselves to be in strong LD ($\text{lod} > 3.0$) with the allele predicted.

Using this tagging procedure in simulated disease association studies as above, we computed power and the number of tag SNPs required. In comparison to pairwise tagging, relative power remained unchanged at 100%, but the number of tag SNPs was reduced by another 26% (YRI), 30% (CEU) and 28% (CHB+JPT; **Fig. 2**). Therefore, by simply removing redundancy from the complete set of SNPs in an efficient haplotype-based manner, we reduced the genotyping burden by 60–77% while maintaining complete power.

Increasing efficiency by relaxing thresholds for tag SNP selection

The tagging strategies described above require that tags be selected to capture perfectly every common site observed in the reference panel. To the extent that this is unaffordable, investigators may be forced to reduce the density of genotyping by relaxing the criteria for tag selection. We examined two possibilities: (i) capturing all common alleles, but at a less stringent r^2 threshold¹⁴, or (ii) capturing only a subset of sites, each at a high r^2 threshold.

Relaxing the threshold from perfect correlation to a slightly lower level ($r^2 \geq 0.8$) substantially decreased the number of tags required (a further decrease of 36% in YRI, 47% in CEU and 55% in CHB+JPT), yet relative power remained almost complete at 96%. This approach can straightforwardly be combined with the multimarker method described above, resulting in even greater efficiency (**Fig. 3a**). Even lower r^2 thresholds resulted in less and less genotyping, but relative power began to decline rapidly. Lowering the r^2 threshold too far (while still requiring that all sites be captured at or above this threshold) can result in performance no better than that of a random collection of SNPs (**Fig. 3a**).

An alternative approach is to rank potential tags according to the number of other SNPs for which they can act as a proxy and then to type the SNPs in this priority order (we call this the ‘best N ’ method). This approach is substantially more efficient than lowering the r^2 threshold: for example, choosing a SNP every 10 kb in this manner (only $\sim 5\%$ of all common SNPs) provides relative power of 77% (YRI), 95% (CEU) and 92% (CHB+JPT). Any such pairwise list can be made more efficient by replacing single-marker tests with appropriate multimarker haplotypes (as above), resulting in the most efficient method of those we examined (**Fig. 3a**).

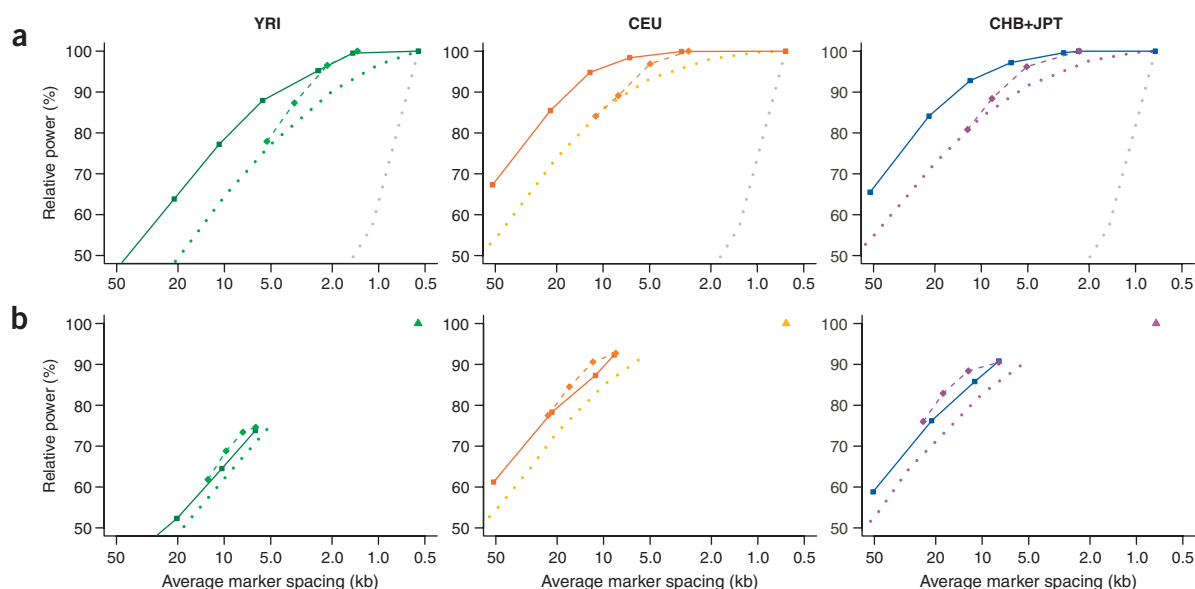


Figure 3 Efficiency and power for various tagging strategies. Relative power to detect associations due to common ($\geq 5\%$) causal alleles is shown as a function of the average spacing of tags picked from (a) complete and (b) incomplete (pseudo phase I HapMap) reference panels. Tags are picked using our multimarker approach by prioritizing best N tags according to number of proxies at $r^2 = 1$ (solid line) and by lowering the r^2 threshold from 1.0 to 0.8, 0.5 and 0.3 (dashed line). Power is also given for random selection of common SNPs tested as single markers (dotted line). In the top panel, expected power is shown for a hypothetical scenario in which there is no LD among SNPs and all tests are independent (gray dotted line); the comparison of this line to the real data shows the gain in efficiency and power offered by the extensive LD in the human genome.

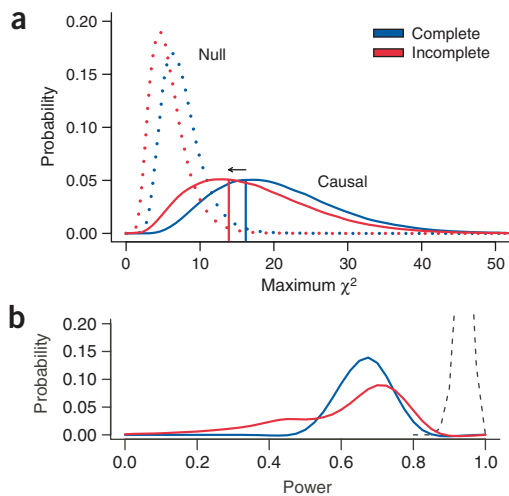


Figure 4 Effect of tagging from an incomplete reference panel on testing burden and power. **(a)** Null (dotted lines) and causal (solid lines) distributions of the test statistic are plotted for two scenarios: tagging from complete (blue) and incomplete (red) reference panels. The causal distribution as well as the region-wide significance threshold are reduced concomitantly when tags are picked from the pseudo phase I HapMap, thereby preserving power. **(b)** Distribution of region-wide power for individual causal alleles is plotted for two scenarios: tagging from complete (blue) and incomplete (red) reference panels. Nominal power is centered around 95% (dotted gray line). Overall power is comparable in both scenarios.

In summary, if a complete reference panel is available, multimer haplotype tests are more efficient than pairwise tests, and prioritizing SNPs on the basis of their LD properties allows impressive reductions in the genotyping burden while maintaining excellent power.

Tags selected from an incomplete reference panel

At present, only incomplete reference panels are available genome-wide^{8,9}. It is therefore important to ask how power and efficiency decline when tags are selected from an incomplete, rather than complete, reference panel. To this end, we created a 'pseudo' 5-kb HapMap by thinning the ENCODE data to achieve the spacing and frequency distribution of phase I HapMap⁸. We selected tags and designed tests using this incomplete resource, evaluating performance in simulated case-control panels where all alleles (not just those from the incomplete HapMap) were allowed to be causal.

We observed two key changes, both predictable. First, a much smaller set of tags was selected for genotyping compared with the set selected when tags were picked using the complete data (**Fig. 3b**). Second, a subset of common variants had no good proxies in the reference panel: 55% (YRI), 26% (CEU) and 28% (CHB+JPT) of all common SNPs were not captured at $r^2 \geq 0.8$, because they were not observed in the pseudo HapMap and were not in LD with any other SNP that was included⁸.

Given these characteristics, it is noteworthy that power was largely undiminished relative

to testing tags chosen from a reference panel of all common sites: tags selected from the pseudo phase I HapMap (pairwise $r^2 \geq 0.8$) provided 91% relative power in CEU (73% in YRI and 89% in CHB+JPT), despite requiring less than one-half of the tags required when tagging from complete data. In absolute terms, a set of best N tags every 10 kb (on average) selected from complete data provided 95% relative power in CEU (77% in YRI and 92% in CHB+JPT), whereas the same density of tags selected from the pseudo phase I HapMap retained 88% power in CEU (64% in YRI and 85% in CHB+JPT).

We also asked whether the power provided by different tagging strategies was similar when using incomplete versus complete reference panels. Whereas the best N strategy outperformed the strategy of lowering the r^2 threshold in complete data, when tagging from the pseudo HapMap (**Fig. 3b**), the two methods performed similarly, although the strategy of lowering the r^2 threshold seemed to have a slight edge.

The impact of LD on tag SNP selection and power

The relatively high power obtained by selecting tags from incomplete reference panels or by using the best N method to trim a complete tag set was somewhat unexpected, because in both cases, the tags did not capture a substantial proportion of putative causal alleles. But the high power is partially explained by the highly variable extent of LD in the human genome and the effects of LD on the power obtained from each statistical test.

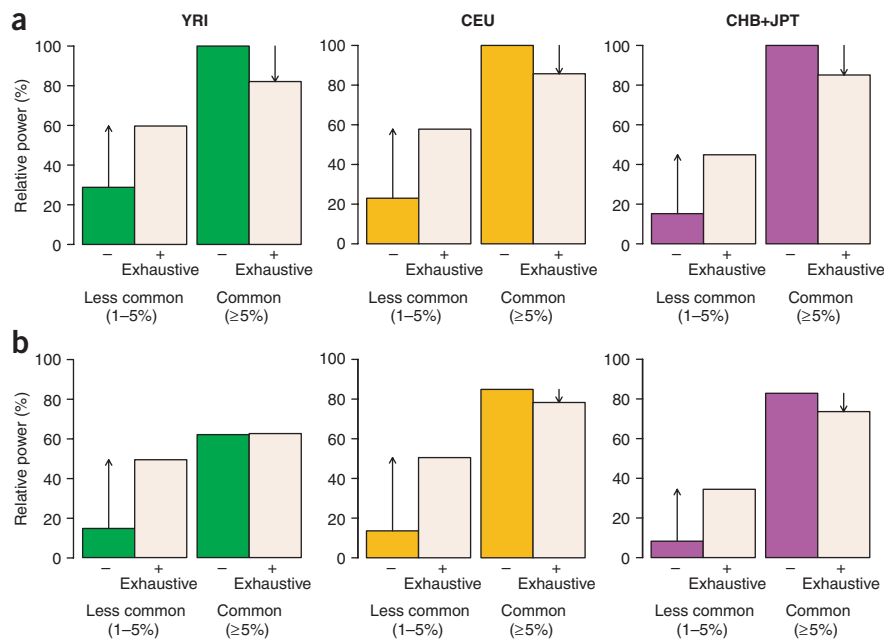


Figure 5 Effect of exhaustive haplotype tests on statistical power. Relative power is given for common ($\geq 5\%$) and less common (1–5%) causal alleles for two scenarios: **(a)** when a nonredundant set of SNPs are used as tags from complete reference panels and **(b)** when tag SNPs (minor allele frequency $\geq 5\%$) are randomly selected every 10 kb from pseudo phase I HapMap (as in ref. 25). Power is computed when each tag SNP is tested for association using single-marker tests (–) and when exhaustive haplotype tests are done on the same data (+). Exhaustive haplotype tests increase power for less common alleles but at a cost of reduced power for common alleles.

The completeness of the reference panel, and the strategy for tagging and testing, affects not only the distribution of the test statistic for causal SNPs but also the significance thresholds under the null distribution. We examined the distribution of the maximum χ^2 test statistic under two scenarios: tags selected from complete and from incomplete reference panels (Fig. 4a). When tags were selected from incomplete data, the causal distribution was shifted toward lower χ^2 values, as expected, because some causal SNPs are not well captured. But in addition, the null distribution shifted to lower thresholds owing to a marked reduction in the number of tests done. Although some alleles were poorly captured and not discovered (most notably those alleles with few proxies; Supplementary Fig. 2 online), the power for most putative causal alleles remained high owing to inclusion of a good proxy for most causal alleles and a less stringent significance threshold for declaring association. Although overall power was similar in both scenarios (Fig. 4b), the mix of causal alleles discovered shifted toward those in LD with many other SNPs, at the cost of discoveries due to SNPs with few proxies.

Put another way, tests that capture many putative causal alleles add the same amount to the multiple testing burden as do independent tests that capture only a single site. The chance of encountering a true association, however, is much greater when many putative causal alleles are examined per test. The single best tag (with most observed proxies) from the incomplete reference panel captures only a small fraction of all sites, but does so at the cost of only a single hypothesis test, and results in relative power that is 15–25% of that obtained by testing all common sites in the region (data not shown). Adding more tags captures a larger fraction of putative causal alleles, and power rises. But the yield of each additional test falls monotonically as it examines a smaller slice of the prior distribution than the test before it.

This simple idea underlies the best N method for tag SNP selection, as it preferentially excludes those SNPs that have no proxies and that offer the least marginal power per hypothesis test. Similarly, an incomplete reference panel (HapMap) has also preferentially (but imperfectly) dropped SNPs with no proxies; such SNPs can be tested for association only if they are included on the HapMap, whereas SNPs with many proxies will almost always be tested, as only one of the proxies needs to be present on HapMap. The best N approach underperforms at sparser densities (in complete and incomplete data; Fig. 3), however, because the set of SNPs with no proxies has been depleted and, therefore, the dropped tags carry with them information about an increasingly larger number of putative causal alleles.

The best N method suffers further when applied to incomplete reference panels, because from such data it is not possible to distinguish which SNPs truly have no proxies and which have proxies that have not yet been observed. Empirically, $\sim 50\%$ of the SNPs on the phase I HapMap have no observed proxies (at $r^2 = 1$)⁸ and are therefore preferentially dropped using the best N method. Of these, a large number have proxies in the complete data, but it is impossible to tell which these are without more complete data. Therefore, where complete data is available (as in selected candidate genes³³), and as denser versions of HapMap become available (such as the pending phase II), the utility of the best N method should increase, particularly for choosing marker densities of more than one SNP per 10 kb.

Exhaustive haplotype tests to detect less common alleles

The above analyses considered only scenarios in which the causal alleles are common. But less common SNPs also influence disease and might be discovered incidentally even if tags are selected and tests designed to capture only common variation. We therefore examined power under the scenario that the frequency of the causal allele is

1–5% with the same 95% nominal power (and thus a larger magnitude of effect). Although power for alleles with this frequency was lower than that for common alleles, it remained substantial: relative power was 29% (YRI), 23% (CEU) and 15% (CHB+JPT).

Exhaustive haplotype testing has been suggested as an approach to capture alleles not observed in the reference panel. This approach tests many or all local haplotypes in the hope that one or more might correspond to an unobserved causal allele²⁵. But this potential benefit comes at the cost of carrying out numerous additional statistical tests, many of which do not correspond to any actual variant.

We first evaluated a scenario in which exhaustive haplotype testing was done on tags picked to capture all common alleles in the complete reference panel ($r^2 = 1.0$) but the universe of causal alleles was limited to those with frequency 1–5%. As described previously²⁵, exhaustive haplotype testing increased relative power: 59% (YRI), 58% (CEU) and 45% (CHB+JPT; Fig. 5a). Therefore, for less common alleles, the benefit of finding a better proxy outweighed the cost of multiple comparisons and resulted in substantial power.

In contrast, when the causal alleles were common ($\geq 5\%$), relative power was reduced by exhaustive haplotype testing to $\sim 85\%$ (Fig. 5a). This penalty was not unexpected: the testing burden was increased with no possibility of true benefit, because all putative causal alleles were already captured.

It seemed more likely that exhaustive haplotype tests might improve power for tags selected from incomplete data or at random. When we selected tags from the incomplete (pseudo phase I HapMap) reference panel or at random at lower densities (one common SNP per 10 kb or 30 kb), exhaustive haplotype tests continued to boost power for less common alleles but did not improve power for common alleles (Fig. 5b and Supplementary Fig. 3 online). We conclude that in empirical genotype data the benefit of exhaustive haplotype tests is real but primarily limited to lower frequency alleles.

Software

The optimal trade-off between power and efficiency depends on the resources available and assumed characteristics of allele frequency and LD for putative causal alleles. Because investigators will want to make their own decisions, we implemented these methods in the web server Tagger and the program Haploview³⁴. The software enables investigators to select tags from empirical data, using single-marker or specified multimarker tests; to rank-order the tags according to proxy count; and to record the statistical tests to be done on these tags (single-marker tests, specified multimarker tests or exhaustive tests). Haploview can carry out association tests based on these selections, including permutation testing. The software also makes it possible to force in or exclude specific sets of SNPs as tags identified on the basis of other considerations, such as the existence of previous data or a working assay; to incorporate genotyping platform design scores to pick tags on the basis of the likelihood of success; to evaluate the coverage with respect to a reference panel (based on r^2) for an existing set of user-specified tags; and to derive specified multimarker tests from a static list of tags to extend coverage with respect to a reference panel (such as HapMap).

DISCUSSION

From our analyses we draw several conclusions. First, specified multimarker tests substantially increase tagging efficiency relative to single-marker approaches, without loss of power. Second, when selecting SNPs from very dense reference panels, a method such as the best N strategy, which ranks SNPs according to the number of proxies they have, allows marked reductions in genotyping with limited loss of

power, substantially outperforming a method based on relaxing r^2 thresholds. Third, sparser sets of tags selected from a pseudo phase I HapMap are almost as powerful as equally sized sets chosen from complete reference panels. Fourth, exhaustive multimarker tests improve power for less common causal alleles but are neutral or reduce power when the causal SNP is common. These relationships hold for each of the different population samples studied by HapMap, although the number and performance of tags varies, as expected, according to the general extent of LD in each sample.

It has become common practice to select tags until a high threshold for the correlation coefficient (often $r^2 \geq 0.8$) is exceeded for all observed sites¹⁴. The use of multimarker tests and prioritization of tags allows for substantial cost reduction with little loss of power. Whether it is worthwhile to take advantage of this trade-off of efficiency for power will be determined by each investigator, depending on the resources available for genotyping, the sample size and power, the perceived cost of a false negative study and the anticipated value of a true positive result.

Whether exhaustive haplotype testing is justified depends on assumptions about the relative balance of rare and common causal variants and the completeness of the reference panel from which tags are picked. Given the current phase I HapMap, there seems to be little cost and evident gain in using the exhaustive haplotype test²⁵. As reference panels become more complete (particularly for less common alleles), however, the balance may shift toward the specified haplotype-based method that limits tests to only those that predict the increasingly complete inventory of putative causal sites.

A limitation of our study is that we did not evaluate whether tags and tests defined in the HapMap samples are transferable across populations. In preliminary analyses, we observed minimal loss of power when tags and tests were transferred to various disease studies (P.L.W.d.B., N. Burt, R. Graham, M.J.D. & D.A., unpublished results), and similar findings have been reported elsewhere^{11,35,36}. Much more work is needed on this topic, and the answer will probably vary depending on the population studied.

Perhaps the most important observation in this study is that SNPs that capture many putative causal alleles have different statistical properties than tests capturing only a single site (at least under the frequentist approach to setting statistical thresholds). An implication of this is that rather than using a universal significance threshold for all tests, power may be increased by a Bayesian approach in which a prior for each test is established as a function of the number of sites captured, integrated over each site's individual likelihood of being causal. Incorporating such ideas into study design may lead to greater efficiency in use of genotyping resources and maximize the yield of discoveries for a given investment in such research.

METHODS

Data sets. We used phased genotype data for ten chromosomal regions, each spanning 500 kb, generated as part of the HapMap ENCODE project. This data set (release 16c.1) was created by genotyping all variable sites observed after resequencing 48 unrelated individuals (as well as any additional SNPs in dbSNP) in the 269 DNA samples used in HapMap (YRI, CEU, CHB and JPT). We combined the CHB and JPT samples for all analyses, yielding three analysis panels: YRI (120 unrelated chromosomes), CEU (120 unrelated chromosomes) and CHB+JPT (178 chromosomes).

Genetic model and simulation of case-control panels. From the ENCODE data, we generated almost 10 million case-control panels to evaluate study-wide power as a function of a number of tagging and testing strategies. We used a multiplicative disease model in which we nominated all nonsingleton SNPs in the complete data to be causal, one by one, reflecting a uniform prior

probability of any of the SNPs contributing to the phenotype. For each causal SNP, we made 250 replicate case-control panels by sampling with replacement from the ENCODE chromosomes to give 1,000 cases and 1,000 controls (4,000 chromosomes in total). The frequency of the causal allele (minor or major chosen at random) in the cases is determined by the genotype relative risk, calibrated so that we obtain 95% nominal power to detect an association with the 1 d.f. χ^2 test (at $P < 0.01$), if that causal SNP was tested directly (**Supplementary Fig. 1**). Thus, all causal SNPs are assigned to have equal nominal power. We also created control-control (null) panels by randomly sampling from the ENCODE chromosomes; we used these to define statistical significance thresholds (**Supplementary Note**).

Reference panels for tag SNP selection. We constructed reference panels at two densities: 'complete' reference panels using all ENCODE data (120 unique chromosomes for YRI and CEU; 178 for CHB+JPT), where complete refers to the ascertainment of common ($\geq 5\%$) variation, and 'incomplete' reference panels, made by thinning the data as follows. To mimic the ascertainment scheme of the 5-kb HapMap (phase I), we randomly picked SNPs present in dbSNP build 121 (excluding 'non-rs' SNPs in HapMap release 16a) for every 5-kb bin until we picked a common (minor allele frequency $\geq 5\%$) SNP (allowing up to three attempts per bin).

Selection of tag SNPs and definition of tests. We developed a computer program called Tagger for selecting tag SNPs and defining tests from a reference panel. Tags can be picked in different ways: (i) greedy pairwise tagging¹⁴, in which alleles of interest are captured by single-marker tests at the prescribed r^2 ; (ii) prioritizing tags (best N) by the number of alleles for which they can serve as a proxy at a given r^2 . In addition, Tagger can carry out an aggressive search to attempt to replace each tag with a specific multimarker predictor (on the basis of the remaining tags) to improve efficiency. This predictor will be accepted only if it can capture the alleles originally captured by that discarded tag at the required r^2 ; otherwise, that tag is considered indispensable. As a result of this 'peel back' approach, we end up with fewer tags that specify a similar (identical if $r^2 = 1$) set of 1 d.f. statistical tests as the original set of single-marker tests. In this study, we allow up to three tags to form a specified multimarker test and limit the search to evaluate at most 10,000 allelic predictors. The maximum allowed physical distance between an allele and a tag was 200 kb. To minimize risk of overfitting, tags in a specified multimarker test are forced to be in strong LD (here defined as lod score > 3) with one another and with the predicted allele.

Region-wide test statistic and power calculations. For every explored tagging and testing scenario, we generated a set of 1 d.f. χ^2 allelic tests. Our region-wide test statistic for association is the maximum of these χ^2 values. The null distribution of the test statistic was generated by carrying out the same allelic tests in the random null panels and used to derive the significance threshold corresponding to a region-wide $P = 0.01$ (a brief discussion on how this compares with explicit permutation testing is presented in **Supplementary Note**). The absolute power to detect association is computed as the fraction of the case-control panels in which the maximal χ^2 test statistic exceeds the significance threshold (when a true association is declared). To normalize results for different strategies, we report power (for both common and rare causal alleles) relative to the power to detect common causal alleles (minor allele frequency $\geq 5\%$) when these are tested directly, averaged over all ten ENCODE regions.

Exhaustive haplotype tests. We carried out exhaustive haplotype tests by enumerating all haplotypes corresponding to adjacent combinations of tags of all sliding windows of a maximum span. We applied this to pairwise tags (selected at $r^2 = 1$ from complete panel) forming haplotypes of up to 25 kb, and 17 and 50 random common markers per region (30 kb and 10 kb average spacing, respectively) from incomplete reference panels forming haplotypes of up to 100 kb. Allelic χ^2 tests were done on these haplotypes as described above.

URLs. The HapMap project website is <http://www.hapmap.org/>. Tagger is available at <http://www.broad.mit.edu/mpg/tagger/>, and Haploview is available at <http://www.broad.mit.edu/mpg/haploview/>. The HapMap ENCODE project website is <http://www.hapmap.org/downloads/encode1.html.en>.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank N. Patterson, E. Lander, J. Hirschhorn and S. Schaffner for discussions; J. Barrett and J. Maller for their implementation of Tagger in Haploview; the Broad Systems Group for technical assistance; and members of the Analysis group of the International HapMap Project for many useful interactions. D.A. is a Charles E. Culpeper Scholar of the Rockefeller Brothers Fund and a Burroughs Wellcome Fund Clinical Scholar in Translational Research. This work was supported by grants from the US National Institutes of Health.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at <http://www.nature.com/naturegenetics/>

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>

- Wang, W.Y., Barratt, B.J., Clayton, D.G. & Todd, J.A. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* **6**, 109–118 (2005).
- Carlson, C.S., Eberle, M.A., Kruglyak, L. & Nickerson, D.A. Mapping complex disease loci in whole-genome association studies. *Nature* **429**, 446–452 (2004).
- Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. & Lander, E.S. High-resolution haplotype structure in the human genome. *Nat. Genet.* **29**, 229–232 (2001).
- Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
- Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
- Johnson, G.C. *et al.* Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29**, 233–237 (2001).
- The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
- The International HapMap Consortium. A haplotype map of the human genome. *Nature* (in the press).
- Hinds, D.A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
- Stram, D.O. *et al.* Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum. Hered.* **55**, 27–36 (2003).
- Weale, M.E. *et al.* Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am. J. Hum. Genet.* **73**, 551–565 (2003).
- Ke, X. & Cardon, L.R. Efficient selective screening of haplotype tag SNPs. *Bioinformatics* **19**, 287–288 (2003).
- Meng, Z., Zaykin, D.V., Xu, C.F., Wagner, M. & Ehm, M.G. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am. J. Hum. Genet.* **73**, 115–130 (2003).
- Carlson, C.S. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106–120 (2004).
- Hu, X., Schrod, S.J., Ross, D.A. & Cargill, M. Selecting tagging SNPs for association studies using power calculations from genotype data. *Hum. Hered.* **57**, 156–170 (2004).
- Halldorsson, B.V. *et al.* Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. *Genome Res.* **14**, 1633–1640 (2004).
- Ao, S.I. *et al.* CLUSTAG: hierarchical clustering and graph methods for selecting tag SNPs. *Bioinformatics* **21**, 1735–1736 (2005).
- Zhang, K. *et al.* HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics* **21**, 131–134 (2005).
- Rinaldo, A. *et al.* Characterization of multilocus linkage disequilibrium. *Genet. Epidemiol.* **28**, 193–206 (2005).
- Schaid, D.J., Rowland, C.M., Tines, D.E., Jacobson, R.M. & Poland, G.A. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* **70**, 425–434 (2002).
- Zaykin, D.V. *et al.* Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum. Hered.* **53**, 79–91 (2002).
- Fan, R. & Knapp, M. Genome association studies of complex diseases by case-control designs. *Am. J. Hum. Genet.* **72**, 850–868 (2003).
- Stram, D.O. *et al.* Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum. Hered.* **55**, 179–190 (2003).
- Chapman, J.M., Cooper, J.D., Todd, J.A. & Clayton, D.G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.* **56**, 18–31 (2003).
- Lin, S., Chakravarti, A. & Cutler, D.J. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat. Genet.* **36**, 1181–1188 (2004).
- Roeder, K., Bacanu, S.A., Sonpar, V., Zhang, X. & Devlin, B. Analysis of single-locus tests to detect gene/disease associations. *Genet. Epidemiol.* **28**, 207–219 (2005).
- Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
- Nyholt, D.R. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am. J. Hum. Genet.* **74**, 765–769 (2004).
- Dudbridge, F. & Koeleman, B.P. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am. J. Hum. Genet.* **75**, 424–435 (2004).
- Wang, W.Y. & Todd, J.A. The usefulness of different density SNP maps for disease association studies of common variants. *Hum. Mol. Genet.* **12**, 3145–3149 (2003).
- Goldstein, D.B., Ahmadi, K.R., Weale, M.E. & Wood, N.W. Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet.* **19**, 615–622 (2003).
- Schaffner, S.F. *et al.* Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* (in the press).
- Crawford, D.C. *et al.* Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am. J. Hum. Genet.* **74**, 610–622 (2004).
- Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
- Nejentsev, S. *et al.* Comparative high-resolution analysis of linkage disequilibrium and tag single nucleotide polymorphisms between populations in the vitamin D receptor gene. *Hum. Mol. Genet.* **13**, 1633–1639 (2004).
- Ahmadi, K.R. *et al.* A single-nucleotide polymorphism tagging set for human drug metabolism and transport. *Nat. Genet.* **37**, 84–89 (2005).