nature genetics

Evaluating and improving power in whole-genome association studies using fixed marker sets

Itsik Pe'er[1,4,5], Paul I W de Bakker[1,2,4,6], Julian Maller[1], Roman Yelensky[1,2,7], David Altshuler[1–6] & Mark J Daly[1,4,5]

Emerging technologies make it possible for the first time to genotype hundreds of thousands of SNPs simultaneously, enabling whole-genome association studies. Using empirical genotype data from the International HapMap Project, we evaluate the extent to which the sets of SNPs contained on three whole-genome genotyping arrays capture common SNPs across the genome, and we find that the majority of common SNPs are well captured by these products either directly or through linkage disequilibrium. We explore analytical strategies that use HapMap data to improve power of association studies conducted with these fixed sets of markers and show that limited inclusion of specific haplotype tests in association analysis can increase the fraction of common variants captured by 25–100%. Finally, we introduce a Bayesian approach to association analysis by weighting the likelihood of each statistical test to reflect the number of putative causal alleles to which it is correlated.

Whole-genome association studies are a comprehensive approach to testing the hypothesis that common alleles contribute to heritable phenotype variation[1–3]. Although neither resequencing every base nor typing all 11 million currently known polymorphic sites in the human genome[4] is yet technically feasible, a practical path to genome-wide association studies has been opened by the introduction of genome-wide SNP arrays[5,6] that type 100,000 to 500,000 SNPs per sample.

Such association studies benefit greatly from linkage disequilibrium (LD)[1,2,7], the correlation between the SNPs on each array and other nearby (untyped) putatively causal alleles[8]. With the completion of the Phase II of HapMap[9], it becomes possible to address two important questions with respect to the use of these arrays. First, to what extent do the fixed set of SNPs on these arrays capture the information about common variation in the human genome[10]? Second, is it possible to devise analytical strategies that make use of HapMap data to increase the chance of discovering a true association?

We evaluate three whole-genome products: the 100K and 500K GeneChip Mapping Sets of Affymetrix[6], and the Sentrix HumanHap300 BeadChip by Illumina[5] (products that contain 116,204, 504,152 and 317,503 SNPs, respectively). Figures for the GeneChip 500K and HumanHap300 products are based on lists of SNPs included on the product (rather than established genotyping performance in laboratories around the world), and thus should be considered preliminary, best-case scenarios. Updated information about evaluations of these and subsequent products are available online.

SNPs included on the Affymetrix products have been preselected primarily on the basis of technical quality and thus represent a quasi-random set of SNPs. In contrast, SNPs on the Illumina product were selected using a pairwise correlation-based algorithm applied to genotype data of HapMap Phase I SNPs in the CEU panel (samples collected by the Centre d'Etude du Polymorphisme Humain (CEPH) from Utah residents with European ancestry)[11].

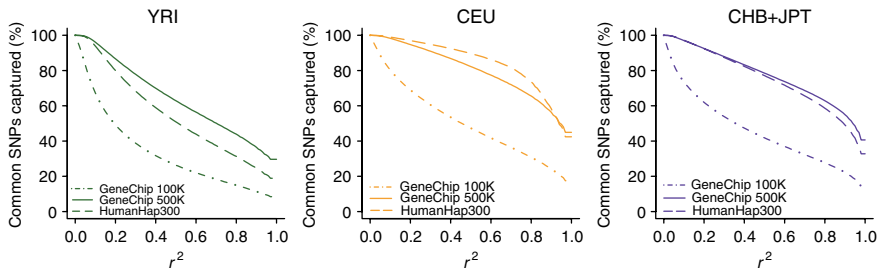

Figure 1 Fraction of common (MAF ≥ 5%) Phase II HapMap SNPs (*y*-axis) captured by array SNPs as a function of the r^2 cutoff (*x*-axis). Data are presented for the GeneChip 100K, GeneChip 500K and HumanHap300 arrays, for each of the three HapMap analysis panels: Yoruba people ascertained in Ibadan, Nigeria (YRI); the CEPH collected samples of European ancestry, ascertained in Utah (CEU); and Han Chinese samples from Beijing with Japanese samples from Tokyo (CHB+JPT).

[1]Center for Human Genetic Research, [2]Department of Molecular Biology and [3]Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [4]Broad Institute of M.I.T. and Harvard, Cambridge, Massachusetts 02142, USA. [5]Department of Medicine and [6]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [7]Harvard-M.I.T. Division of Health Sciences and Technology, Cambridge, Massachusetts 02139, USA. Correspondence should be addressed to M.J.D. (mjdaly@chgr.mgh.harvard.edu) and D.A. (altshul@broad.mit.edu).

Received 21 February; accepted 2 May; published online 21 May 2006; doi:10.1038/ng1816

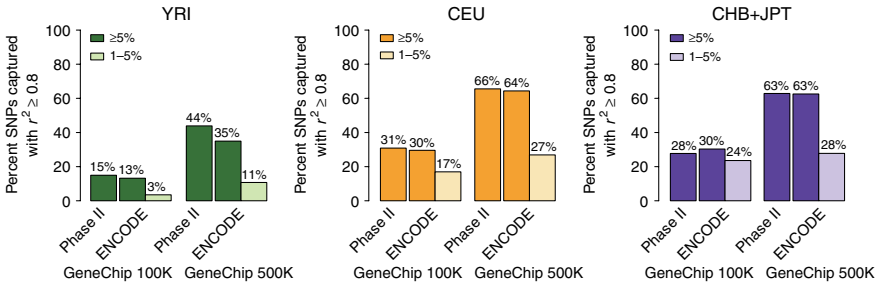

Figure 2 Fraction of SNPs (*y*-axis) captured by SNPs on GeneChip 100K and 500K arrays at $r^2 \geq 0.8$ in the three HapMap panels: YRI, CEU and CHB+JPT. Data are presented for common SNPs (dark bars) as observed in HapMap Phase II and ENCODE, and for less common (MAF 1–5%) SNPs (light bars) as observed in ENCODE. As ENCODE data do not fully represent SNPs from the latter category, but rather include only a partial set of such SNPs that happened to have been discovered (and tend to be more common), results presented here should be considered as upper bounds for the ability to capture the complete set of alleles of frequencies 1–5%.

Ideally, evaluation of each marker set would involve measuring the extent to which it is correlated with every putative causal common allele along the genome. Although complete polymorphism data do not yet exist to support such an analysis, all three array SNP sets have been typed in the HapMap reference samples of 270 individuals from four population samples[9]. These panels therefore allow, in principle, evaluation of correlation in two data sets: the ENCODE data of ten regions spanning 5 Mb, with essentially complete ascertainment for alleles with frequency ≥5% (refs. 12,13), and the genome-wide Phase II HapMap, which includes roughly 3.9 million SNPs successfully typed to date. We therefore evaluate the GeneChip 100K and 500K arrays vis-à-vis ENCODE and evaluate all three arrays on the Phase II HapMap data.

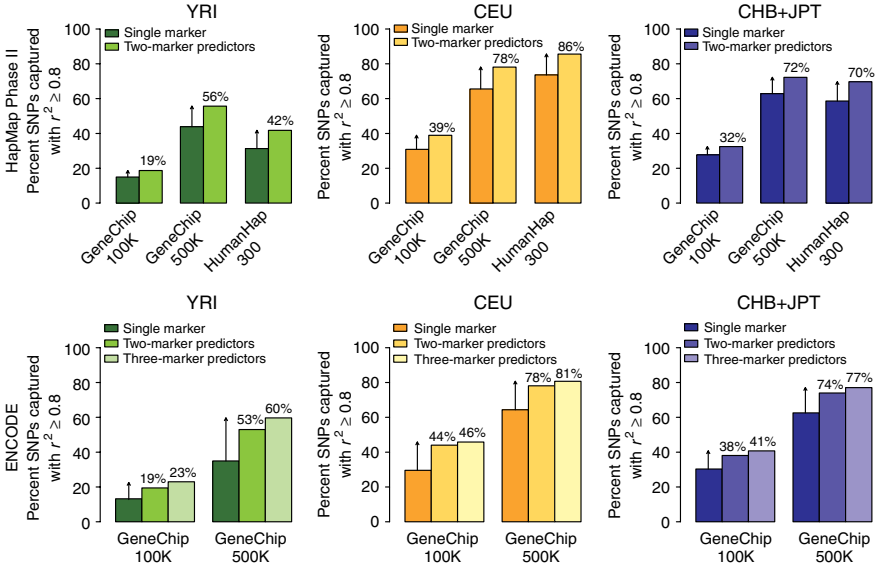

Figure 3 Fraction of common SNPs (*y*-axis) captured by single-array SNPs versus multimarker predictors in three HapMap panels (YRI, CEU and CHB+JPT). Data are presented for HapMap Phase II (top) as well as ENCODE (bottom). For Phase II data, we evaluated the GeneChip 100K, GeneChip 500K and HumanHap300 arrays with two-marker predictors. For ENCODE, we evaluated the GeneChip arrays with two- or three-marker predictors. As SNP selection for the HumanHap300 product is based on LD information from Phase I HapMap data (including ENCODE), evaluation using this data set would be biased upward and therefore is omitted. We report results only for common SNPs in order to minimize risk of overfitting in the multimarker predictors and thus overestimating the ability to capture rare alleles.

A full exploration of the utility of a SNP set involves estimating the power to detect association under many study design and disease scenarios[14]. A simpler, study-independent measure of utility is the square of the correlation coefficient (r^2) between any observed marker and a putative causal allele[15]. This metric is interpretable as the expected drop in non-centrality of an association test statistic under specified conditions[16], and it has become one standard for evaluating performance of marker sets[17–20].

Figure 1 shows the correlation between common SNPs in the Phase II data (that is, SNPs with minor allele frequency (MAF) ≥ 5%) and markers on the whole-genome arrays (see the **Fig. 1** legend for details). The fraction of SNPs captured is a function of the threshold correlation coefficient required for tag SNP selection. For example, in the CEU panel, 45% of all common Phase II SNPs are captured by the GeneChip 500K array at r^2 of 1 (that is, no loss of power compared with testing the putative causal SNP directly), whereas 62% of common SNPs are captured at r^2 of 0.8 and 80% with $r^2 \geq 0.5$ (that is, highly significant correlations to untyped alleles but with modest loss of power in association settings). As expected, SNPs on the array capture a smaller proportion of variants in the most genetically diverse panel, YRI, than are captured in the CEU and CHB+JPT panels, in which the fractions of SNPs captured are higher and similar to one another.

Figure 2 examines correlations of SNPs in the more fully ascertained ENCODE regions for the GeneChip arrays. This cross-validates the results of common Phase II SNPs and allows examination of a substantial, yet incomplete, set of SNPs with frequency 1–5%. The representation of the latter set of SNPs is limited and biased by the scope of SNP discovery efforts, which tend to miss the rarer alleles. The examined set of SNPs therefore demonstrates only an upper bound on the ability of the arrays to capture low-frequency alleles, which is much poorer than corresponding ability for common ones[21]. This highlights the focus of the array content at common variants, where association studies are most powerful to detect (subtle) genetic effects[22]. Comprehensive scans for rare causal alleles will require other sets of markers, more involved analysis methods[23,24] and, where possible, complete resequencing.

Even though the majority of common variants is captured by the current generation of genome-wide arrays, there is a substantial component of common variation not highly correlated to a SNP on each array. We set out to analytically improve the ability to capture common variants using only the SNPs on these arrays and knowledge of LD in available HapMap data. Here we describe an approach in which HapMap data is used to detect correlations between specific combinations of alleles for SNPs on each array (called

multimarker predictors[20]) and a putatively causal allele previously uncaptured. We and others[17–20,25] have elsewhere introduced this concept in the context of tag SNP selection, avoiding the typing of certain SNPs to improve typing efficiency while maintaining study power. In the context of fixed-content SNP genotyping products, we propose to use specific multimarker predictors of untyped SNPs (inferred from the HapMap) as tests of association, thereby increasing study power without performing additional genotyping.

We observe that multimarker predictors based on combinations of alleles of two or three SNPs can capture (at $r^2 \geq 0.8$) an additional 9–25% of SNPs in ENCODE or HapMap Phase II (**Fig. 3**). Notably, using these specific tests (listed online; see Methods), the HumanHap300 and GeneChip 500K arrays gain the ability to capture 80–86% of common alleles in the CEU population with this high level of correlation. These tests also facilitate pooling association results from studies that have used different arrays, through combined predictions of the same SNPs. This gain in power is achieved without additional genotyping and thus permits more comprehensive association studies with current products, at no extra cost.

A possible concern is the potential of overfitting based on HapMap relationships involving limited sample sizes (120 chromosomes for CEU and YRI; 180 chromosomes for CHB+JPT). Mathematically, however, the chance of a highly correlated ($r^2 \geq 0.8$) common variant in this sample size is much smaller ($<10^{-12}$) than the space of predictors searched for each SNP. We verified this empirically by developing multimarker predictors to unlinked SNPs: we never observed spurious correlations of $r^2 > 0.35$ in HapMap data. Although for rare alleles, overfitting is indeed an issue using the HapMap sample sizes, we are confident that relationships at thresholds such as $r^2 > 0.5$ involving common SNPs are robust and reliable.

These results suggest that in situations in which direct typing of a common causal SNP would be successful, use of one of these genotyping arrays will often provide the opportunity to detect that association as well, through LD[20]. However, when any additional testing (such as the addition of the multimarker tests) is performed, the benefits of capturing more variation need to be evaluated against the statistical cost of performing additional hypothesis testing. This is because addition of statistical tests could, in principle, lead to a reduction in power by requiring increased statistical significance thresholds to maintain constant type I error rates (or, conversely, allowing substantially more false positives if statistical thresholds are unchanged). This is because addition of statistical tests could, in principle, lead to a reduction in power by requiring increased statistical significance thresholds to maintain constant type I error rates (or, conversely, allowing substantially more false positives if statistical thresholds are unchanged).

This tradeoff is of particular relevance to multimarker predictors, as they capture on average fewer untyped SNPs than do single SNPs. That is, we observe that statistical tests based on the genotype of a SNP on the array have more proxies on average in HapMap Phase II than do statistical tests based on two- and three-marker haplotype predictors (3.85 versus 1.55 putative causal alleles captured, respectively, on the GeneChip 500K array in the CEU panel). At the extreme, testing all observed allele combinations[26] rather than only the SNPs and specified multimarker predictors might not pay off, as the marked increase in degrees of freedom[18,25] results in only a tiny increase in the fraction (3% in CEU) of common SNPs captured[9]. Adding many tests while increasing information capture by a small amount can result in a loss of power for association to common alleles[18]. Indeed, a recent detailed simulation study[20] shows an increase in power for common causal variant detection when these specific multimarker tests are added but a slight reduction in power when all possible haplotypes are considered.

Next, we consider a Bayesian strategy to tests all alleles without suffering from an increased burden of multiple testing. The standard, frequentist strategy for genome-wide association studies[8] assigns a one- or two-degree-of-freedom score to each variant tested and searches for *P*-values deemed significant. While *P*-values speak to the degree to which observed data is unexpected under the null (that is, no association) hypothesis, external information may be quite relevant to the alternative hypothesis (that is, that the tested or a nearby correlated variant is truly causal). Intuitively, not all tests are created equal—hypothesis tests that capture the genotypic variance at many SNP sites, or tests that correspond to known functional alterations, may rightly be considered more likely *a priori* to be true positives than those hypothesis tests that capture only a single variant site (of unknown functional significance). This highly relevant information is not customarily considered *a priori* in a formal fashion (although it is often discussed in a post-hoc manner). Analysis of the HapMap data makes it possible to incorporate such information up front in association analysis. Specifically, we define prior probabilities based on the identities and number of putative causal alleles captured by each allelic hypothesis test. Having assigned to each allele a prior probability of causality, we can evaluate the *a posteriori* likelihood of association given the data (see Methods).

We demonstrate this framework using one objective, simple and universal hypothesis used in simulation studies[20,26]: namely, that each common SNP in the genome is equally likely to be causal. The *a priori* likelihood of association to each marker on the array is therefore proportional to the number of SNPs it captures. The number of variant sites captured by each hypothesis test is highly variable, as even very large clusters of correlated SNPs may be represented by a single SNP, whereas other SNPs capture only themselves. We show by simulated association studies that incorporation of such prior probabilities (see Methods) modestly but consistently (and statistically significantly) improves power to detect association as compared with a frequentist framework. For example, association testing to 100 SNPs, chosen either randomly or by LD tagging, is improved by 4% by this approach (**Supplementary Fig. 1** online). Moreover, the value of this approach will only increase as genomic annotation improves the estimate of the prior probability of each variant site in the genome being causal. Individual investigators can tailor analysis based on their own views of how to weight SNPs that are coding[27], associated with variation in gene expression, under a compelling linkage peak[28] or in genes whose function is tied to a particular pathway.

The simultaneous emergence of genome-wide genotyping arrays and comprehensive, deeply ascertained SNP data from HapMap provides for the first time a toolkit to evaluate association between common genetic variation and disease throughout the genome. We find that current products capture a sizeable portion of genomic variation, and we describe methods to use the HapMap data for testing additional non-array SNPs *in silico* without further genotyping. Finally, we have developed a framework to prioritize the tested SNPs based on external information provided by HapMap and, potentially, additional genomic annotation. Such methods should help enable systematic and more powerful evaluation of the contribution of common alleles to complex phenotypes.

METHODS

Data sets. We used the phased ENCODE data from HapMap (release 16c.1). We also used genotype data from Phase II HapMap, merged these with the genotype data generated by the GeneChip 500K array, ported the data to

National Center for Biotechnology Information (NCBI) build 35 (University of California, Santa Cruz (UCSC) hg17) and subsequently phased the final data using the expectation-maximization (EM) algorithm[29].

Choosing multimarker predictors. For every array product, we have specified a set of haplotype tests based on HapMap using Tagger[20]. For every SNP that is not typed on the array, we aim to find the allelic test (predictor) with the highest r^2 to that SNP, exploiting the knowledge of which SNPs are present on the array. The predictors are identified by performing an aggressive search among combinations of two or three SNPs (on the array), evaluating the r^2 between the generated haplotypes and the allele we want to capture. Although many of the untyped SNPs are captured by high pairwise correlation to a SNP on the array, a substantial fraction of the (common) SNPs is not. The multimarker predictors for all three arrays evaluated here can be found on our website.

Simulating case-control panels. Our simulation framework follows a recently published protocol[20]. Briefly, the phased ENCODE chromosomes ($n = 120$ from unrelated individuals in CEU) were resampled to create 1,000 cases and 1,000 controls (4,000 chromosomes in total). For controls, resampling was uniform. For cases, we designated one SNP to be causal. For this causal SNP, we calculated an effect size (and corresponding allele frequency in the cases) such that if it were to be the only SNP tested, power would be 95% to detect it at a nominal P value of 0.01. In terms of relative risk, the simulated effect size was therefore larger for rare alleles (**Supplementary Fig. 2** online). We created 250 case-control panels for each causal SNP, where we allowed, at random, either allele of a given SNP to be causal. We repeated this for all common SNPs in a region and for all ten ENCODE regions separately (a total of nearly 10,000 SNPs). We also generated 250,000 null panels (without a causal SNP) for evaluation of the null distribution.

Power calculations. Power is defined as the fraction of the simulated case-control panels in which the test statistic exceeds the significance threshold (when an association can be declared), averaged over all ten ENCODE regions. We use the maximum of the 2×2 χ^2 comparison over all allelic tests (single-marker tests and, optionally, the specified multimarker tests) as the region-wide test statistic. The significance threshold is derived by performing the same allelic tests from the null panels (to achieve a region-wide corrected P value of 0.01). The absolute power to detect association at $P < 0.01$ after multiple testing correction is 68%, if all common SNPs are evaluated. Power remains >90% of this figure when the best tags (with most proxies) are selected at a density of one tag per 5 kb, if these tests are given uniform weights[20].

Derivation of weights for allelic tests. Suppose the set of m putative causal alleles is $A = \{a_1, \ldots, a_m\}$. Denote by $C[a_i, I]$ the count of the allele a_i in a set I of individuals. Let I_1, I_0 be sets of cases and controls of sizes N_1 and N_0, respectively. Define the normalized difference statistic

$$Z(a_i) = \frac{\frac{C[a_i,I_1]}{N_1} - \frac{C[a_i,I_0]}{N_0}}{\sqrt{(C[a_i,I_1]+C[a_i,I_0])\left(1 - \frac{C[a_i,I_1]+C[a_i,I_0]}{N_1+N_0}\right)}}$$

Suppose further that the set of n tests (single- or multimarker predictors) used to capture these alleles is $T = \{t_1, \ldots, t_n\}$, and extend the definition of the count operator [] and the statistic Z to these tests.

The null hypothesis is simple: $Z(t_1), \ldots, Z(t_n)$ are all standard normal variables (also known as z-scores). In contrast, the alternative hypothesis is complex: it states that a causal allele is chosen out of A according to some prior distribution D: $A \rightarrow [0,1]$ (where $D(a_i)$ denotes the probability of a_i to be chosen as causal), and given that choice, all tests that are correlated with a_c are normally distributed with means greater than zero. More specifically, let μ_c be the effect size for the causal allele a_c, represented in terms of mean offset of $Z(a_c)$ from the origin. For each test t_j, let $r_{c,j}$ denote its correlation coefficient to a_c. Hence, if a_c is causal, $Z(t_j)$ is normally distributed with mean $\mu_c r_{c,j}$.

In this study, we denote the normal probability density function (p.d.f.) and cumulative density function (c.d.f.) by ϕ and Φ, respectively. We use the simulation assumption[20] that $\mu_c = \Phi(0.95) + \Phi(0.99) \approx 3.97$. We use Haploview[30] to compute the matrix $R = [r_{c,j}]_{m \times n}$ of correlation coefficients between all tests and all alleles and transform it into a matrix $W = [w_{i,j}]_{m \times n}$ where $w_{i,j}$ is the probability that, given a_i is causal, it will be detected by t_j; that is, the top-scoring test for a_i is t_j, and it is above the null signal. Formally, if $r_{i,j} = 0$, we set $w_{i,j}$ to zero as well. Otherwise, to compute $w_{i,j}$ we integrate over the real signal, $Z(a_i)$, given which we can write the score distributions of the current test t_j and the scores in needs to exceed: the null signal, as well as any true signal by some other test $t_{j'}$ correlated to a_i. We approximate such tests as being dependent through a_i only. We can thus express all relevant probabilities as functions of $Z(a_i)$, as follows:

$$\begin{aligned} w_{i,j} &\approx \int_{z=-\infty}^{\infty} P(Z(t_j) > null | Z(a_i) = z) \\ &\quad \bullet \left(\prod_{j' | r_{i,j} \neq 0, j \neq j'} P(Z(t_j) > Z(t_{j'}) | Z(a_i) = z) \right) \bullet \phi(z - \mu_i) dz \approx \\ &= \int_{z=-\infty}^{\infty} P_{null}(r_{i,j} \cdot z) \bullet \left(\prod_{j' | r_{i,j'} \neq 0, j \neq j'} \Phi\left(\frac{z(r_j - r_{j'})}{\sqrt{2 - r_j^2 - r_{j'}^2}} \right) \right) \\ &\quad \bullet \phi(z - \mu_i) dz \end{aligned}$$

where *null* represents the region maximum score in a null panel and $P_{null}(z)$ is the empirically derived c.d.f. of this maximum. We can now write the likelihood ratio test for a data set with the maximum-scoring test t_j achieving an observed score z_j:

$$\frac{\Pr[Z(t_j) = z_j | H_1]}{\Pr[Z(t_j) = z_j | H_0]} = \left[\sum_{i=1}^{m} (D(a_i) \bullet w_{i,j}) \right] \bullet P - value(z_j)$$

We thus compute a weight factor $W_j = \sum_{i=1}^{m} (D(a_i) \bullet w_{i,j})$ for each test t_j employed, and use that to prioritize all P values.

URLs. International HapMap Project ENCODE: http://www.hapmap.org/downloads/encode1.html.en; whole-genome association products: http://www.broad.mit.edu/mpg/wga-products; International HapMap Project: http://www.hapmap.org/; Tagger: http://www.broad.mit.edu/mpg/tagger/; Haploview: http://www.broad.mit.edu/mpg/haploview/.

Updated information about evaluations of the products described in this paper are available online http://www.broad.mit.edu/mpg/wga-products; this site will be updated as new products become available.

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

ACKNOWLEDGMENTS

We acknowledge Affymetrix, Inc. and Illumina, Inc. for sharing product data. We also thank Affymetrix, Inc. for making public genotype data of the HapMap samples generated by the GeneChip Mapping 500K Array.

AUTHORS' CONTRIBUTIONS

D.A. and M.J.D. jointly supervised this work.

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/

1. Devlin, B. & Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322 (1995).
2. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
3. Collins, F.S., Brooks, L.D. & Chakravarti, A.A. DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8**, 1229–1231 (1998).
4. Wheeler, D.L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **34**, D173–D180 (2006).
5. Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G. & Chee, M.S. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.* **37**, 549–554 (2005).

6. Matsuzaki, H. *et al.* Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods* **1**, 109–111 (2004).
7. Reich, D.E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
8. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
9. Altshuler, D. *et al.* A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
10. Kruglyak, L. Power tools for human genetics. *Nat. Genet.* **37**, 1299–1300 (2005).
11. 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).
12. Kruglyak, L. & Nickerson, D.A. Variation is the spice of life. *Nat. Genet.* **27**, 234–236 (2001).
13. Pe'er, I. *et al.* Biases and reconciliation in estimates of linkage disequilibrium in the human genome. *Am. J. Hum. Genet.* **78**, 588–603 (2006).
14. Purcell, S., Cherny, S.S. & Sham, P.C. Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**, 149–150 (2003).
15. Pritchard, J.K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
16. Sham, P.C., Cherny, S.S., Purcell, S. & Hewitt, J.K. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am. J. Hum. Genet.* **66**, 1616–1630 (2000).
17. 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).
18. 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).
19. 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).
20. de Bakker, P.I. *et al.* Efficiency and power in genetic association studies. *Nat. Genet.* **37**, 1217–1223 (2005).
21. Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H. & Nielsen, R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15**, 1496–1502 (2005).
22. Pritchard, J.K. & Cox, N.J. The allelic architecture of human disease genes: common disease-common variant or not? *Hum. Mol. Genet.* **11**, 2417–2423 (2002).
23. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in *PCSK9*. *Nat. Genet.* **37**, 161–165 (2005).
24. Cohen, J.C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869–872 (2004).
25. 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).
26. 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).
27. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33**(Suppl.), 228–237 (2003).
28. Roeder, K., Bacanu, S.A., Wasserman, L. & Devlin, B. Using linkage genome scans to improve power of association in genome scans. *Am. J. Hum. Genet.* **78**, 243–252 (2006).
29. Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**, 921–927 (1995).
30. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).