nature genetics

Transferability of tag SNPs in genetic association studies in multiple populations

Paul I W de Bakker[1–4,16], Noël P Burtt[1,16], Robert R Graham[1–4,16], Candace Guiducci[1], Roman Yelensky[1–3,5], Jared A Drake[1,6], Todd Bersaglieri[1,6], Kathryn L Penney[7], Johannah Butler[1,6], Stanton Young[2,3], Robert C Onofrio[1], Helen N Lyon[1,6], Daniel O Stram[8], Christopher A Haiman[8], Matthew L Freedman[1,9], Xiaofeng Zhu[10], Richard Cooper[10], Leif Groop[11,12], Laurence N Kolonel[13], Brian E Henderson[8], Mark J Daly[1,2,14], Joel N Hirschhorn[1,4,6] & David Altshuler[1–4,14,15]

A general question for linkage disequilibrium–based association studies is how power to detect an association is compromised when tag SNPs are chosen from data in one population sample and then deployed in another sample. Specifically, it is important to know how well tags picked from the HapMap DNA samples capture the variation in other samples. To address this, we collected dense data uniformly across the four HapMap population samples and eleven other population samples. We picked tag SNPs using genotype data we collected in the HapMap samples and then evaluated the effective coverage of these tags in comparison to the entire set of common variants observed in the other samples. We simulated case-control association studies in the non-HapMap samples under a disease model of modest risk, and we observed little loss in power. These results demonstrate that the HapMap DNA samples can be used to select tags for genome-wide association studies in many samples around the world.

The International HapMap Project provides empirical genotype data for >3 million SNPs in a limited sample of 270 individuals from four populations[1]. There are two fundamental questions with regard to a dense reference panel such as HapMap. First, to what extent is power compromised when tags are selected from incomplete genotype data in the reference panel? Second, how is power affected when tags are selected from a reference panel but then genotyped in another population sample? Previously, we addressed the first question by evaluating the quantitative relationship between marker density and power in simulated association studies using HapMap ENCODE data with near-complete ascertainment of common variation[2]. Here, we characterize the extent to which tag SNPs picked from HapMap DNA samples are transferable across different population samples.

To this end, we have collected dense genotype data uniformly across HapMap and non-HapMap population samples. As part of the Multiethnic Cohort (MEC) Study[3,4], we compiled a list of genes in the steroid hormone and growth factor pathways. In each of these genes, we selected a dense set of SNPs from the public dbSNP database[5] and augmented this set by SNP discovery through exon resequencing in 190 individuals with breast and prostate cancer from five different ethnic groups (**Supplementary Table 1** online). In total, we attempted genotyping for 3,302 SNPs in >1,000 DNA samples from 15 population samples (**Table 1** and **Supplementary Table 2** online). We kept all SNPs that were successfully genotyped in all population samples and that were polymorphic in at least one (**Supplementary Table 3** online). The final data set contained 1,679 SNPs across 25 genes with a total span of 2.6 Mb and an average marker density of 1 SNP per 1.6 kb (**Supplementary Table 4** online).

To assess the transferability of tags picked from HapMap samples for association studies in other population samples, two relevant measures are (i) the distribution of the correlation (r^2) between the allelic tests (based on the tags) and all 'untyped' variants (that is, SNPs not selected as tags) present in these samples and (ii) how this translates into study-wide power to detect an association under a specified disease model. We prefer these measures to comparisons based on differences in linkage disequilibrium (LD) structure, haplotype diversity or allele frequencies, as the question of immediate interest is the impact of any such differences on the power in the disease study.

[1]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Seven Cambridge Center, Cambridge, Massachusetts, 02142, USA. [2]Center for Human Genetic Research and [3]Department of Molecular Biology, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114-2790, USA. [4]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [5]Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts, USA. [6]Divisions of Endocrinology and Genetics, Program in Genomics, Children's Hospital, Boston, Massachusetts 02115, USA. [7]Harvard School of Public Health, Boston, Massachusetts 02115, USA. [8]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90089, USA. [9]Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. [10]Department of Preventive Medicine and Epidemiology, Loyola University, Maywood, Illinois 60153, USA. [11]Department of Clinical Sciences, University Hospital, Lund University, Malmö S-20502, Sweden. [12]Department of Medicine, Helsinki University, Helsinki, Finland. [13]Cancer Research Center, University of Hawaii, Honolulu, Hawaii 96813, USA. [14]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA. [15]Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [16]These authors contributed equally to this work. Correspondence should be addressed to D.A. (altshuler@molbio.mgh.harvard.edu).

Received 23 March; accepted 12 September; published online 22 October 2006; doi:10.1038/ng1899

Table 1 Population samples included in this study

Samples and origin	Abbreviation	Number of chromosomes in final data set[a]
Yoruba from Ibadan, Nigeria, International HapMap Project	YRI	120
Yoruba from Ibadan, Nigeria, Human Genome Diversity Project	HGDP-YRI	50
African Americans from Los Angeles, Multiethnic Cohort	MEC-AA	138
African Americans from Chicago	MAY	96
Utah, European ancestry from CEPH, International HapMap Project	CEU	104
Utah, European ancestry from CEPH (other)	CEPH-EXT	248
Self-described 'whites' from Hawaii, from Multiethnic Cohort	MEC-W	136
Botnia, Finland	BOT	116
Han Chinese from Beijing, International HapMap Project	CHB	88
Han Chinese from Beijing, Human Genome Diversity Project	HGDP-CHB	80
Japanese from Tokyo, International HapMap Project	JPT	88
Japanese from Tokyo, Human Genome Diversity Project	HGDP-JPT	62
Japanese from Hawaii and Los Angeles, Multiethnic Cohort	MEC-J	136
Native Hawaiians from Hawaii, Multiethnic Cohort	MEC-H	138
Latinos from Los Angeles, Multiethnic Cohort	MEC-L	138

[a]After QC and data filtering.

Using only the genotype data collected in each HapMap panel (YRI, CEU, CHB and JPT), we selected tags until every SNP observed with $\geq$5% allele frequency in that panel was captured with a pairwise $r^2 \geq 0.8$ by at least one tag. By definition, these tags capture all (including 'untyped') common SNPs ($\geq$5%) at a maximum $r^2 \geq 0.8$ in that HapMap panel. The mean maximum r^2 between the tags and the 'untyped' SNPs was 0.93–0.96 (**Table 2**), and many 'untyped' SNPs had a perfect proxy (maximum $r^2 = 1$) (**Fig. 1a–d**).

Before considering transferability across population samples, it is crucial to measure the effect of transferability to a second, independent sample from the same population. Specifically, we expect to see statistical fluctuation in allele frequencies around the 5% threshold for tag SNP selection. For example, SNPs with an estimated allele frequency just below the 5% threshold will not be targeted during tag SNP selection (and may not be captured) but may well have an allele frequency above this threshold in a second sample[6]. Conversely, SNPs with an estimated allele frequency just above the 5% threshold will be captured by a tag but may fall below the threshold in a second sample and therefore may not be included in the assessment. Furthermore, there is fluctuation in the estimated r^2 for pairs of SNPs in independent samples of limited size: SNPs captured with $r^2 \geq 0.8$ by a tag in one sample may be captured with $r^2 < 0.8$ in a second sample (and vice versa) owing to random fluctuations in the chromosomes chosen for each sample. These effects are a natural consequence of sampling variability and employing strict allele frequency and r^2 thresholds.

We characterized the extent of sampling variation in the HapMap reference panels by evaluating the coverage of common SNPs in independent samples drawn from the same population. The vast majority of the 'untyped' common SNPs were still captured with a maximum $r^2 \geq 0.8$: 71% in unrelated individuals from Ibadan, Nigeria also used in the Human Genome Diversity Project (HGDP-YRI), 89% in parent-offspring trios from Utah (from the Centre d'Etude du Polymorphisme Humain) with northern and western European ancestry (CEPH-EXT), 79% in unrelated Han Chinese from Beijing from the HGDP (HGDP-CHB) and 78% in unrelated Japanese from Tokyo from the HGDP (HGDP-JPT) (**Table 2**). In fact, most SNPs were captured with a maximum $r^2 \geq 0.5$ (**Fig. 1e–h**). The observed loss is interpreted as statistical fluctuation caused solely by drawing independent samples of limited size from the same underlying population, resulting in modest r^2 overestimation during tag SNP selection.

A small fraction of 'untyped' SNPs, however, were not well captured: between 2% and 11% of 'untyped' SNPs had a maximum $r^2 < 0.5$ in the second, independent sample from the same population (**Fig. 1e–h**). Upon closer inspection, these poorly captured SNPs typically had a lower allele frequency (**Fig. 2**). All SNPs with a maximum $r^2 < 0.5$ in CEPH-EXT had a frequency below 5% in the HapMap CEU panel (a few were monomorphic) and were consequently missed, as no tags were explicitly picked to capture them. Thus, the observed loss is due to the fluctuations in the allele frequency estimates, not to differences in LD structure. We observed this thresholding effect, to a lesser degree, in HGDP-YRI, HGDP-CHB and HGDP-JPT. Some 'untyped' SNPs were captured poorly even though they were observed at $\geq$5% frequency in the reference panel, reflecting a decrease in LD between the tags and these SNPs.

Table 2 Coverage of common variation in the population samples by the tags picked from the HapMap samples

			All common SNPs		'Untyped' common SNPs	
Reference panel (HapMap samples)	Number of picked tags	Population sample	Mean maximum r^2	Percentage with maximum $r^2 \geq 0.8$	Mean maximum r^2	Percentage with maximum $r^2 \geq 0.8$
YRI	724	YRI	0.97	100%	0.93	100%
YRI	724	HGDP-YRI	0.92	87%	0.82	71%
YRI	724	MEC-AA	0.87	79%	0.73	56%
YRI	724	MAY	0.87	80%	0.73	56%
CEU	470	CEU	0.97	100%	0.95	100%
CEU	470	CEPH-EXT	0.95	93%	0.91	89%
CEU	470	MEC-W	0.92	86%	0.87	78%
CEU	470	BOT	0.93	89%	0.89	83%
CHB	388	CHB	0.97	100%	0.95	100%
CHB	388	HGDP-CHB	0.91	86%	0.87	79%
JPT	415	JPT	0.97	100%	0.96	100%
JPT	415	HGDP-JPT	0.92	85%	0.88	78%
JPT	415	MEC-J	0.94	91%	0.91	86%

Coverage is expressed as the mean maximum r^2 and the percentage with maximum $r^2 \geq 0.8$ for all SNPs (including tags) and 'untyped' SNPs with frequency $\geq$5% in each population sample.

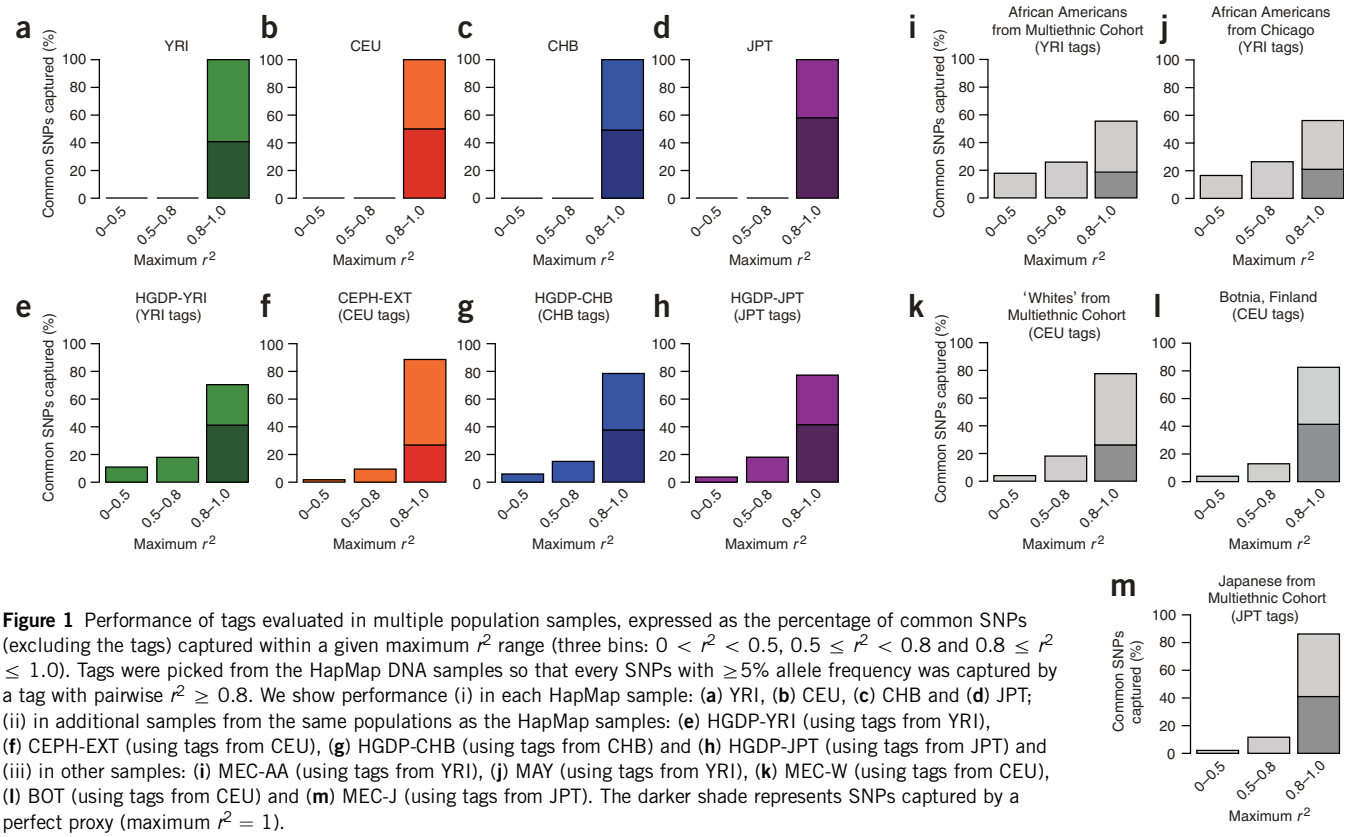

Figure 1 Performance of tags evaluated in multiple population samples, expressed as the percentage of common SNPs (excluding the tags) captured within a given maximum r^2 range (three bins: $0 < r^2 < 0.5$, $0.5 \leq r^2 < 0.8$ and $0.8 \leq r^2 \leq 1.0$). Tags were picked from the HapMap DNA samples so that every SNPs with ≥5% allele frequency was captured by a tag with pairwise $r^2 \geq 0.8$. We show performance (i) in each HapMap sample: (**a**) YRI, (**b**) CEU, (**c**) CHB and (**d**) JPT; (ii) in additional samples from the same populations as the HapMap samples: (**e**) HGDP-YRI (using tags from YRI), (**f**) CEPH-EXT (using tags from CEU), (**g**) HGDP-CHB (using tags from CHB) and (**h**) HGDP-JPT (using tags from JPT) and (iii) in other samples: (**i**) MEC-AA (using tags from YRI), (**j**) MAY (using tags from YRI), (**k**) MEC-W (using tags from CEU), (**l**) BOT (using tags from CEU) and (**m**) MEC-J (using tags from JPT). The darker shade represents SNPs captured by a perfect proxy (maximum $r^2 = 1$).

Having assessed the transferability within samples of the same population, we next examined transferability to samples with similar continental ancestry as the CEU and JPT HapMap panels but sampled from different populations. Using the tags picked from HapMap CEU samples, we evaluated the coverage of common variants in self-described 'white' individuals from Hawaii (MEC-W) and in individuals from the Botnia region of Finland (BOT). The performance of the CEU tags in these non-HapMap samples was very similar to that in CEPH-EXT (**Fig. 1k–l**). The mean maximum r^2 for 'untyped' SNPs was 0.87–0.89 (**Table 2**). In both samples, only 4% of 'untyped' variants were captured with a maximum $r^2 < 0.5$. We also evaluated the coverage in Japanese samples from Hawaii and Los Angeles (MEC-J) for tags picked from HapMap JPT samples (**Fig. 1m**). Performance was similar: 86% of 'untyped' variants were captured with a maximum $r^2 \geq 0.8$ with a mean maximum r^2 of 0.91 (**Table 2**), and only 2% were captured with a maximum $r^2 < 0.5$. Thus, in these samples there is little loss in coverage beyond that observed in independent samples from the same population.

We next evaluated the performance of tags picked from the YRI HapMap samples in African American samples from Los Angeles (MEC-AA) and from Chicago (MAY). Of all 'untyped' common variants, 56% were captured with a maximum $r^2 \geq 0.8$, with a mean maximum r^2 of 0.73 (**Table 2**). A comparatively larger fraction of SNPs (18%) was captured with a maximum $r^2 < 0.5$ (**Fig. 1i–j**). We did not find this surprising in light of estimates that African Americans have 80%–85% African ancestry[7]. A tagging strategy that takes into account the combined African and European ancestry of these samples would be expected to improve coverage, as we will demonstrate below.

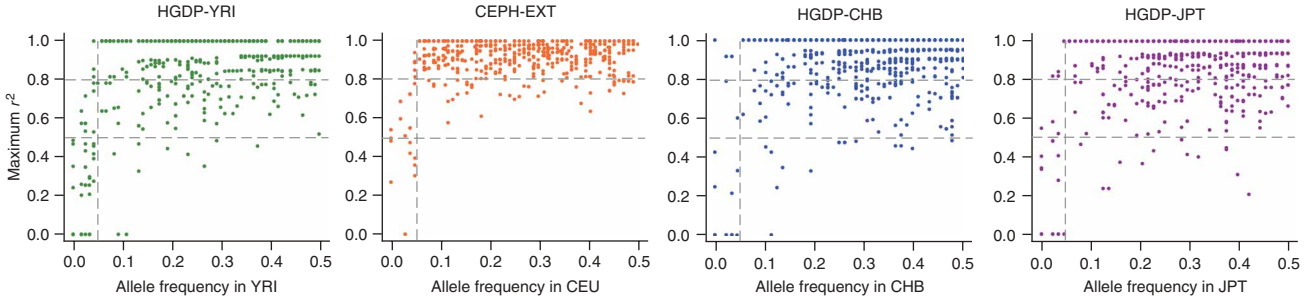

Figure 2 The relationship between the allele frequency observed in the HapMap reference panel (from which tags are picked) and the maximum r^2 between the tags and all 'untyped' SNPs with ≥5% frequency in HGDP-YRI, CEPH-EXT, HGDP-CHB and HGDP-JPT. Tags were picked from each HapMap reference panel to capture all SNPs with ≥5% frequency in that panel (threshold indicated by vertical lines). No explicit attempt was made to capture SNPs with <5% frequency in the reference panel. SNPs that are poorly captured (maximum $r^2 < 0.5$) tend to have lower allele frequency. The gray horizontal lines separate the three r^2 bins used in **Figure 1**.

Table 3 Relative power in simulated case-control association studies in non-HapMap populations

Reference panel (HapMap samples)	Number of picked tags	Simulated case-control panel	Relative power for all causal alleles (%)	Relative power for 'untyped' causal alleles (%)
YRI	724	HGDP-YRI	95	87
YRI	724	MEC-AA	92	81
YRI	724	MAY	92	81
CEU	470	CEPH-EXT	97	95
CEU	470	MEC-W	96	91
CEU	470	BOT	96	89
CEU	470	MEC-H	94	89
CEU	470	MEC-L	92	83
CHB	388	HGDP-CHB	96	92
JPT	415	HGDP-JPT	96	92
JPT	415	MEC-J	96	92
Cosmopolitan	885	MEC-AA	96	90
Cosmopolitan	885	MAY	96	89
Cosmopolitan	885	MEC-H	99	96
Cosmopolitan	885	MEC-L	97	94

Tags were picked from a reference panel (HapMap sample) to capture all SNPs observed with ≥5% frequency in that panel with a pairwise maximum $r^2 \geq 0.8$. The relative power is the power to detect causal alleles (SNPs with ≥5% frequency in the non-HapMap population sample) in comparison with the observed power when all causal alleles are tested directly (no tagging). Power is given for all common SNPs as well as for the subset of common SNPs that were not picked as tags ('untyped'). 'Cosmopolitan' refers to the set of tags that captures all common SNPs across all four HapMap populations.

Although these results are encouraging, we wanted to obtain a more direct estimate of statistical power in a disease study. Because of the correlated nature of dense SNP data and the number of statistical tests, it is not straightforward to estimate power directly from the r^2 distribution. Thus, we simulated case-control association studies for each non-HapMap population sample using a recently described procedure[2]: for each non-HapMap sample, we nominated each common SNP with ≥5% allele frequency as 'causal' in turn, and we generated a large number of simulated case-control panels. We evaluated power by performing the association tests (based on the same tags selected from the HapMap samples as above) in these case-control panels and counting the number of panels in which we were able to detect an association at a gene-wide corrected P value of 0.01, averaged over all 25 genes.

Testing all common SNPs in the simulated case-control panels resulted in an absolute power of 82% in HGDP-YRI, 84% in CEPH-EXT and HGDP-CHB and 85% in HGDP-JPT. This is substantially higher than the power in HapMap ENCODE data under the identical genetic model (60% in YRI and 68% in CEU and CHB+JPT)[2]. This is due to an overall reduction in the multiple testing burden, because the genes in the present data set are smaller and have less complete ascertainment than the 500-kb regions of the HapMap ENCODE project. We report the power to detect all common causal alleles, as well as the power to detect 'untyped' common variants as a conservative estimator, and we express both relative to the power obtained by testing all common SNPs in the case-control sample (as if we had genotyped all common SNPs directly).

In independent samples from the same populations as the HapMap samples, power was 87%–95% for the 'untyped' common SNPs relative to the power obtained by testing all common SNPs in those samples (**Table 3**). Relative power decreased by only ~5% in MEC-W and BOT (using CEU tags) and remained unchanged in MEC-J (using JPT tags). Performance was lower in the African American MEC-AA and MAY samples: relative power was 81% using tags picked in YRI. (For the sake of comparison, if we used the CEU tags alone, relative power dropped to 63%.)

We attempted to improve the power for the African American samples by picking tags from all four HapMap population samples combined (rather than picking tags from YRI only). At an additional genotyping cost (22% more tags), this 'cosmopolitan' tagging approach increased the relative power to 89%–90% for the 'untyped' common SNPs in both African American samples (**Table 3**). This demonstrates that tags from the HapMap populations are able to provide good power in these samples. It is likely that tag SNP selection could be made more efficient by incorporating knowledge about the ancestry of samples. Power did not deteriorate when tags were picked from YRI and CEU only (excluding CHB and JPT), but it did result in fewer tags. This is not unexpected: tags from CHB and JPT that are not redundant with tags from YRI and CEU include SNPs that are unique to these two East Asian populations.

For some population samples like native Hawaiians (MEC-H) and Latinos (MEC-L) from the MEC, there is no obvious choice of HapMap reference panel from which to pick tags. Nevertheless, when we used CEU in our initial attempt, relative power was 89% for 'untyped' variants in MEC-H, and power in MEC-L was only slightly worse (**Table 3**). The cosmopolitan tags improved relative power to 94%–96% in both population samples, albeit at a greater genotyping cost.

Recently, we introduced specified multimarker (haplotype) tests as a means to improve upon pairwise tagging in terms of genotyping efficiency without sacrificing power[2]. In this approach, specific haplotypes act as effective surrogates for single tag SNPs. This keeps the multiple testing burden constant while decreasing the number of tags (but requiring greater genotyping quality and completeness). When we used this 'aggressive' tagging approach to capture all common SNPs with $r^2 \geq 0.8$, the genotyping burden was reduced by 15%–23% compared with pairwise tagging. Power in the simulated case-control association studies remained essentially unchanged with this more efficient tagging approach (data not shown). Hence, this multimarker strategy is robust to transferability, at least for the DNA samples tested here. A notable implication of this result is that specified multimarker tests inferred from a dense reference panel (such as HapMap) can act as effective predictors for some untyped SNPs. We have recently demonstrated that this approach can provide a substantial boost in the coverage of commercially available whole-genome products[8].

Our results are broadly consistent with other assessments of tag transferability[9–20]. Although LD structure is known to vary between population samples[12,21–24], we have demonstrated empirically that a standard tagging approach in the four HapMap population samples can capture common variation effectively in many other independent samples. Most of the observed loss in coverage and power results from fluctuations in allele frequency and r^2 estimates owing to sampling

variation. We conclude that tag SNPs selected from the HapMap DNA samples are likely to provide good power to study the role of common polymorphisms in complex traits in many sample collections.

METHODS

DNA samples. We collected genotype data in the following DNA samples: 30 parent-offspring trios from the Yoruba people in Ibadan, Nigeria (YRI), 27 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) used the International HapMap Project[1]; 25 unrelated individuals from Ibadan, Nigeria (HGDP-YRI), 40 unrelated Han Chinese from Beijing, China (HGDP-CHB) and 31 unrelated Japanese from Tokyo (HGDP-JPT) from the Human Genome Diversity Project[25,26]; 62 trios from Utah with northern and western European ancestry, from the CEPH collection (CEPH-EXT); 70 self-described African American (MEC-AA), 69 self-described native Hawaiian (MEC-H), 70 self-described Japanese (MEC-J), 70 self-described Latino (MEC-L) and 70 self-described 'white' (MEC-W) samples from the Multiethnic Cohort Study conducted in Hawaii and California (mainly Los Angeles); 30 trios from Botnia, Finland (BOT) and 48 unrelated African Americans from Chicago (MAY). These studies were approved by the Human Subject Institutional Review Boards at the respective institutions, and informed consent was obtained from all subjects.

SNP discovery. We performed exon resequencing in 95 cases of advanced breast cancer and 95 cases of advanced prostate cancer from the Multiethnic Cohort Study. These are 19 samples from each of the five populations represented in the Multiethnic Cohort (see above) that do not overlap with the samples used to collect genotype data). Summary statistics are given in **Supplementary Table 1**. Resequencing in these samples was specifically performed to enrich for common functional (missense, splice site, UTR) variants not present in dbSNP at the time (we note that this project started before the International HapMap Project). Of the 542 SNPs that were discovered by resequencing in the 25 genes, 157 were present already in dbSNP and not attempted for validation. Of the remaining novel SNPs, 355 were processed for further validation (primers were designed successfully), and 240 were validated by genotyping in the resequencing panel. There are only 14 SNPs that are nonsynonymous, with a frequency of $>1\%$ in the entire resequencing panel ($n = 190$), and only seven nonsynonymous SNPs had a frequency of $>1\%$ in the Multiethnic Cohort panel ($n = 349$). All SNPs used in the analysis were validated.

SNP genotyping. A dense set of SNPs was selected for genotyping from two sources: (i) SNPs discovered by resequencing that were not in dbSNP (version 117) and that were located in exons or UTRs and, subsequently, (ii) SNPs from dbSNP (version 119) and Celera databases prioritizing 'double-hit' and missense SNPs. Genotyping was performed to generate an initial map of roughly evenly spaced SNPs in the African American MEC samples to classify regions according to their degree of LD and to provide a guide for further genotyping. SNP density was preferentially increased in regions of low(er) LD as inferred from the initial map. In total, we attempted 3,302 SNP assays in 1,029 samples using the Sequenom MassArray and Illumina GoldenGate platforms (**Supplementary Table 2**). Concordance between the Sequenom and Illumina platforms was 98.2% (12,927 out of 13,170) for 863 markers typed in 16 identical samples. In the 15 population samples, on average, 84% (2,774) of the attempted assays passed quality control filters, defined as genotyping completeness $>90\%$, no more than one concordance error, no more than one Mendel inheritance error and $P > 0.001$ for the Hardy-Weinberg test (**Supplementary Table 3**). This resulted in a working set of 1,842 SNPs that passed quality control in all 15 population samples, including 1,679 SNPs that are polymorphic in at least one population sample (1,473 SNPs with $\geq 5\%$ frequency; **Supplementary Table 4**). All genotype data were phased using the program PHASE 2.1.1 (ref. 27) to produce phased chromosomes that were used in all analyses[26]. For the purposes of estimating (high) r^2 values between SNPs, the impact of potential phasing errors is expected to be minimal[28].

Simulation of case-control association studies. For all non-HapMap samples, we simulated case-control panels of each gene to evaluate the power to detect an association. We used a multiplicative genetic model in which we designated all common SNPs, one by one, to be 'causal'. For each causal SNP, we made case-control panels by sampling with replacement chromosomes from the phased data to give 1,000 cases and 1,000 controls (4,000 chromosomes in total). As a function of the allele frequency of the designated 'causal' allele, we set the genotype relative risk to obtain a constant 95% power at a nominal P of 0.01 using a 2×2 χ^2 test. We generated 75 replicate case-control panels per causal SNP, and all SNPs have an equal chance of being causal. We also generated 75,000 control-control (null) panels by sampling chromosomes at random from the phased data. These null panels have no causal SNP and were used to derive gene-wide significance thresholds.

Tag SNP selection. We used the program Tagger to derive a set of tag SNPs from the HapMap reference panel such that each allele that satisfies the allele frequency threshold is captured at the given r^2 threshold either by a single tag (pairwise tagging)[29] or by a specified multimarker (haplotype) test ('aggressive' tagging)[2]. We noticed that the efficiency gain afforded by aggressive tagging was less than that observed in previous analyses of the HapMap-ENCODE data[1,2]. This is likely due to the fact that this study focuses on gene regions of $\sim$100 kb in size, compared with 500-kb ENCODE regions where tagging can benefit from long-range LD.

Cosmopolitan tagging. We have implemented a 'cosmopolitan' tagging approach that maximizes in a greedy fashion, for every additional tag, the total number of captured alleles (at the user-defined threshold: here, $r^2 \geq 0.8$) in multiple reference panels. To illustrate, our approach will pick a SNP with more proxies in multiple reference panels (such as CEU and YRI) as a tag in favor of a SNP with only few proxies in a single panel (such as CEU). This procedure does not directly combine the (phased) genotype data for multiple panels (which could inflate or deflate correlations between markers), but rather evaluates multiple panels in parallel. This is similar in spirit to the TagIT[13] and MultiPop-TagSelect methods[30].

Power calculations. We evaluated power by performing the allelic tests (based on the selected tags) in the simulated null panels and the case-control panels. We derived significance thresholds from the null panels that correspond to a gene-wide corrected P of 0.01. The power is the fraction of case-control panels in which we observed a test statistic greater than the significance threshold. We report the power relative to that obtained by testing all common SNPs in that non-HapMap sample directly, averaged over all 25 genes.

URLs. Genotype data and SNP summary statistics can be found at http://www.uscnorris.com/Core/DocManager/DocumentList.aspx?CID=13. Tagger can be found at http://www.broad.mit.edu/mpg/tagger/.

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

ACKNOWLEDGMENTS

We thank M. Egyud for sharing unpublished results and all members of the collaborative Multiethnic Cohort Study and the Analysis group of the International HapMap Consortium for useful discussions. We acknowledge the support of NIH grants CA63464 and CA098758 (to B.E.H.), HL074166 (to X.Z.), CA54281 (to L.N.K.) and DK067288 (to H.N.L.); a March of Dimes grant (6-FY04-61, to J.N.H.) and a Charles E. Culpeper Scholarship of the Rockefeller Brothers Fund and a Burroughs Wellcome Fund Clinical Scholarship in Translational Research (both to D.A.).

AUTHOR CONTRIBUTIONS

H.N.L., X.Z., R.C., L.G., C.A.H., L.N.K., B.E.H. provided DNA samples; N.P.B. coordinated resequencing with R.C.O. and S.Y.; N.P.B., R.R.G., C.G., J.B., K.L.P. prepared DNA samples, designed and performed genotyping experiments; P.d.B., N.P.B., R.R.G., R.Y., J.A.D. and T.B. performed the analyses; P.d.B. wrote the paper, with contributions from N.P.B. and R.R.G.; M.L.F., C.A.H., D.O.S. and H.N.L. gave feedback and helped with revisions and M.J.D., J.N.H. and D.A. jointly directed the project.

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. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
2. de Bakker, P.I.W. *et al.* Efficiency and power in genetic association studies. *Nat. Genet.* **37**, 1217–1223 (2005).
3. Kolonel, L.N. *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.* **151**, 346–357 (2000).
4. Kolonel, L.N., Altshuler, D. & Henderson, B.E. The multiethnic cohort study: exploring genes, lifestyle and cancer risk. *Nat. Rev. Cancer* **4**, 519–527 (2004).
5. Wheeler, D.L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **34**, D173–D180 (2006).
6. Zeggini, E. *et al.* An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. *Nat. Genet.* **37**, 1320–1322 (2005).
7. Parra, E.J. *et al.* Estimating African American admixture proportions by use of population-specific alleles. *Am. J. Hum. Genet.* **63**, 1839–1851 (1998).
8. Pe'er, I. *et al.* Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.* **38**, 663–667 (2006).
9. 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).
10. 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).
11. Ke, X. *et al.* Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Hum. Mol. Genet.* **13**, 2557–2565 (2004).
12. Mueller, J.C. *et al.* Linkage disequilibrium patterns and tagSNP transferability among European populations. *Am. J. Hum. Genet.* **76**, 387–398 (2005).
13. Ahmadi, K.R. *et al.* A single-nucleotide polymorphism tagging set for human drug metabolism and transport. *Nat. Genet.* **37**, 84–89 (2005).
14. Ramirez-Soriano, A. *et al.* Haplotype tagging efficiency in worldwide populations in CTLA4 gene. *Genes Immun.* **6**, 646–657 (2005).
15. Ribas, G. *et al.* Evaluating HapMap SNP data transferability in a large-scale genotyping project involving 175 cancer-associated genes. *Hum. Genet.* **118**, 669–679 (2006).
16. Stankovich, J. *et al.* On the utility of data from the International HapMap Project for Australian association studies. *Hum. Genet.* **119**, 220–222 (2006).
17. Huang, W. *et al.* Linkage disequilibrium sharing and haplotype-tagged SNP portability between populations. *Proc. Natl. Acad. Sci. USA* **103**, 1418–1421 (2006).
18. Gonzalez-Neira, A. *et al.* The portability of tagSNPs across populations: a worldwide survey. *Genome Res.* **16**, 323–330 (2006).
19. Montpetit, A. *et al.* An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. *PLoS Genet.* **2**, e27 (2006).
20. Smith, E.M. *et al.* Comparison of linkage disequilibrium patterns between the HapMap CEPH samples and a family-based cohort of Northern European descent. *Genomics* published online 19 May 2006 (doi:10.1016/j.ygeno.2006.04.004).
21. Shifman, S., Kuypers, J., Kokoris, M., Yakir, B. & Darvasi, A. Linkage disequilibrium patterns of the human genome across populations. *Hum. Mol. Genet.* **12**, 771–776 (2003).
22. Beaty, T.H. *et al.* Haplotype diversity in 11 candidate genes across four populations. *Genetics* **171**, 259–267 (2005).
23. Evans, D.M. & Cardon, L.R. A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *Am. J. Hum. Genet.* **76**, 681–687 (2005).
24. Sawyer, S.L. *et al.* Linkage disequilibrium patterns vary substantially among populations. *Eur. J. Hum. Genet.* **13**, 677–686 (2005).
25. Cann, H.M. *et al.* A human genome diversity cell line panel. *Science* **296**, 261–262 (2002).
26. Rosenberg, N.A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
27. Stephens, M. & Donnelly, P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73**, 1162–1169 (2003).
28. Marchini, J. *et al.* A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* **78**, 437–450 (2006).
29. 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).
30. Howie, B.N., Carlson, C.S., Rieder, M.J. & Nickerson, D.A. Efficient selection of tagging single-nucleotide polymorphisms in multiple populations. *Hum. Genet.* **120**, 58–68 (2006).