Published online 28 August 2007

Nucleic Acids Research, 2007, Vol. 35, No. 17 ***e113***
doi:10.1093/nar/gkm621

Efficacy assessment of SNP sets for genome-wide disease association studies

Andreas Wollstein[1,2], Alexander Herrmann[2,3], Michael Wittig[2], Michael Nothnagel[4], Andre Franke[2], Peter Nürnberg[1], Stefan Schreiber[2], Michael Krawczak[4] and Jochen Hampe[2,3,*]

[1]Cologne Center for Genomics, Cologne, [2]Institute of Clinical Molecular Biology, Christian-Albrechts University, [3]Ist Department of Medicine and [4]Institute of Medical Informatics and Statistics, Christian-Albrechts University, University Hospital Schleswig-Holstein Campus Kiel, Kiel, Germany

Received May 31, 2007; Revised and Accepted July 30, 2007

ABSTRACT

The power of a genome-wide disease association study depends critically upon the properties of the marker set used, particularly the number and physical spacing of markers, and the level of inter-marker association due to linkage disequilibrium. Extending our previously devised theoretical framework for the entropy-based selection of genetic markers, we have developed a local measure of the efficacy of a marker set, relative to including a maximally polymorphic single nucleotide polymorphism (SNP) at the map position of interest. Using this quantitative criterion, we evaluated five currently available SNP sets, namely Affymetrix 100K and 500K, and Illumina 100K, 300K and 550K in the CEU, YRI and JPT + CHB HapMap populations. At 50% relative efficacy, the commercial marker sets cover between 19 and 68% of the human genome, depending upon the population under study. An optimal technology-independent 500K marker set constructed from HapMap for Caucasians, in contrast, would achieve 73% coverage at the same relative efficacy.

INTRODUCTION

Genome-wide association studies with large sets of single nucleotide polymorphisms (SNP) (1) are a new option for mapping the genetic variants underlying complex human diseases. However, the power and cost-effectiveness of such studies depends critically upon the properties of the SNP sets used. Consequently, the choice between one of the commercially available marker panels and the construction of a new set is of strong practical significance. No objective criteria other than descriptive measures (e.g. marker number) have so far been used to compare the utility of genome-wide marker sets. More importantly, any sensible assessment of a marker panel requires that recent discoveries about the biology of meiotic recombination are appropriately taken into account (2–4). For example, it has been shown (2) that the 'geodesy' of the human genetic map is fairly homogenous above the centi-Morgan level, but that the correlation between physical and genetic distance is weak at a finer scale, due to rapidly evolving recombination hotspots. Consequently, SNP selection strategies that are based upon the assumption of static linkage disequilibrium (LD) blocks, or that merely employ pairwise LD, may result in sub-optimal marker sets.

The utility of a marker set for disease association analysis is determined by a number of factors, including marker number, informativity and spacing, in addition to the local level of LD. In practice, genotyping technologies may pose serious restrictions upon the usability of an individual SNP, irrespective of whether its inclusion might be desirable or not. If such limitations can be ignored, however, then the utility of a marker set should ideally be evaluated by a criterion that:

(i) allows the assessment of the coverage of a genomic region in a single quantity,
(ii) is computationally practicable,
(iii) is applicable to the limited genotype information typically available for large marker sets and
(iv) draws upon a theoretical framework that allows meaningful interpretation of the numerical results.

Shannon entropy (5) is a well-established mathematical concept for assessing the utility of genetic markers. We have recently devised an entropy-based SNP selection approach (6) that can in principle be adapted to a genome-wide setting. Furthermore, the methodology facilitates estimation of the relative, region-specific efficacy of a given marker set by τ, a quantity that approximates to the relative sample size required to map a causative variant

*To whom correspondence should be addressed. Tel: +49 431 597 1246; Fax: +49 431 597 1842; Email: J.Hampe@1med.uni-kiel.de

at a given map position, compared to including a maximally polymorphic SNP at the same position (see Methods section). We calculated τ across the genome using publicly available genotype data for HapMap (Phase 2, built 35) (7) and for the five commercial marker sets of Affymetrix (8) (100K and 500K) and Illumina (9) (100K, 300K and 550K). The results were compared to an 'ideal' SNP set constructed from HapMap via entropy-based marker selection.

METHODS

Estimation of inverse swept radii

Parameter ε, which denotes the inverse of the swept radius, was used as a local measure of LD strength (10,11) and was estimated from HapMap genotype data on the basis of all markers with a minor allele frequency $\geq 10\%$. To this end, pairwise haplotype frequencies were estimated from the genotype data using an EM algorithm. Then, the pairwise allelic association was quantified as

$$\rho = \frac{\det|\mathrm{P}|}{Q(1-R)}, \tag{1}$$

where P is the haplotype frequency matrix $(p_{ij})_{i,j=1\ldots2}$, $Q = p_{11} + p_{12}$ and $R = p_{11} + p_{21}$ (10), and where $\det|\mathrm{P}|$ denotes the determinant of P. Marker-specific ε values were estimated by a log-linear regression analysis of ρ and the physical distance to all other markers X_i in a 500 kb window surrounding the marker Y of interest (12), i.e. by fitting model $\log(\rho) = -\varepsilon \cdot |x_i - y|$ to marker locations x_i and y.

Here and in the following, we assumed that the population of interest was characterized by monophyletic inheritance and by a lack of association between unlinked loci, a simplification of the original model of LD decay that was justified by empirical observations made for autosomal markers in Europe and the US (11).

At inter-marker positions $z_1 < z < z_2$, $\varepsilon(z)$ was estimated by linear interpolation, i.e.

$$\varepsilon(z) = \frac{\varepsilon(z_2) - \varepsilon(z_1)}{z_2 - z_1} \cdot (z - z_1) + \varepsilon(z_1). \tag{2}$$

Entropy-based SNP selection

We have previously devised a method for assessing the utility of marker sets for disease association studies (6), based upon Shannon entropy (5). In brief, for a locus X with k alleles of frequency p_i ($i = 1\ldots k$), entropy $H(X)$ is defined as

$$H(X) = -\sum_{i=1}^{k} p_i \log_2 p_i \tag{3}$$

For the purposes of disease association analysis, a genomic region is assumed to be covered by markers $X_1, \ldots, X_n$ at map positions $x_1 < \ldots < x_n$. Then, the problem of SNP selection reduces to deciding, on the basis of existing genotype or haplotype data, which single marker out of some additional markers $Y_1, \ldots, Y_m$ to include in order to maximize the mapping utility of the extended panel. Without loss of generality, it can be assumed that this choice is confined to maximizing the utility of the marker set in a given interval, centred at map position z. A utility score $\kappa(Y{:}X, z)$ is then constructed that reflects the benefit, with respect to mapping a disease gene at position z, of adding Y to a single marker X,

$$\kappa(Y : X,z) = \mathrm{e}^{-e\cdot\varepsilon(z)\cdot|y-z|} \cdot H(Y|X). \tag{4}$$

Here, $H(Y|X) = H(X,Y) - H(X)$ denotes the conditional entropy of Y given X. The quantity in formula (4) can be calculated directly from pairwise haplotype frequencies, known swept radii and known marker locations. The best marker to include into the existing marker panel $X_1, \ldots, X_n$ is then chosen according to

$$Y_{\max} = \arg\left[\max_{j=1..m} \min_{i=1..n} \kappa(Y_j : X_i,z)\right]. \tag{5}$$

Application to genome-wide marker panels

Application of the above-mentioned framework to large-scale genome-wide data sets poses additional computational problems since the comprehensive evaluation of all pairwise haplotype frequencies, as required by formulas (4) and (5), is no longer feasible. Thus, $\kappa(Y{:}X, z)$ was replaced by

$$\kappa(Y : z) = \mathrm{e}^{-e\cdot\varepsilon(z)\cdot|y-z|} \cdot H(Y) \tag{6}$$

when the distance between Y and z exceeded $3/\varepsilon(z)$ (11). In this way, the number of pairwise haplotype frequency estimations was limited and the computing time scaled linearly (instead of quadratically) with marker number. Formula (6) was also used for selecting the first few markers on a given chromosome, successively breaking the chromosome down into shorter intervals by applying formula (6) to the corresponding interval centers. Marker selection according to formula (4) commenced for an interval when it was shorter than three times the internal median swept radius.

Evaluation of genome-wide marker sets

Following Hampe *et al.* (6), we define criterion $\tau(z)$ for the local evaluation of a marker set around map position z as

$$\tau(z) = \mathrm{e}^{-2\cdot\varepsilon(z)\cdot|x-z|} \cdot \frac{q_{X(z)}}{1 - q_{X(z)}} \tag{7}$$

where $q_{X(z)}$ is the minor allele frequency of that marker, $X(z)$, that maximizes the right-hand side of formula (7) [note that $\tau(z)$ is similar, but not equivalent, to $1 - \kappa_{\min}(z)$ as defined in the original paper (6)]. Since

$$\rho = \mathrm{e}^{-\varepsilon(z)\cdot|x-z|} \tag{8}$$

equals the predicted allelic association (11) between X and a maximally informative biallelic marker Z at map position z, it follows that

$$\tau(z) = \rho^2 \cdot \frac{q_X}{1 - q_X}. \tag{9}$$

On the other hand, the number n of individuals required to detect association ρ between X and Z at significance level α and with power $1-\beta$ is approximately equal to

$$n(q,\rho) \sim \frac{\left[z_{1-\alpha/2} + z_{1-\beta}\right]^2}{2} \cdot \frac{1-q}{q\rho^2} \quad (10)$$

where $z_{1-\alpha/2}$ and z_β are the respective quantiles of the Gaussian distribution (for a detailed derivation of formula (10), see Appendix). For any two marker sets A and B, let $\tau_A(z)$ and $\tau_B(z)$ be the τ values obtained with respect to the same location z. Then,

$$\frac{\tau_A(z)}{\tau_B(z)} = \left(\frac{\rho_A}{\rho_B}\right)^2 \cdot \frac{q_A \cdot (1-q_B)}{q_B \cdot (1-q_A)} \sim \frac{n(q_B,\rho_B)}{n(q_A,\rho_A)} \quad (11)$$

which implies that $\tau(z)$ is a good approximation of the relative efficacy of a marker set, measured by the inverse of the sample size required to map a maximally informative SNP at position z.

Computer implementation

The methodology described above has been implemented into a suite of JAVA programs interacting with a MySQL relational database for the storage of genotypes and intermediate results. Since the HapMap data set was the most exhaustive one, calculation of swept radii was based upon these markers and genotypes. The software is available as a web service under http://www.ikmb.uni-kiel.de/snpselection/.

SNP data sources and genotyping

Caucasian genotype data for HapMap (Phase II, built 35), Affymetrix 100K and 500K were retrieved from the respective web sites (www.hapmap.org, www.affymetrix.com). The marker identities of the Illumina 100K, 300K and 500K sets were retrieved from the Illumina website (www.illumina.com); the corresponding genotypes were taken from HapMap or from the Illumina website.

RESULTS

Quantity τ measures the relative efficacy of a given marker set to map a causal variant at a specified map position, compared to including a maximally polymorphic SNP at the very same position (see Methods section). Therefore, $\tau = 1$ corresponds to full local efficacy of a marker panel whereas $\tau = 0$ indicates that no information can be extracted locally. For the purpose of comparing different marker sets, τ was calculated here at 10 kb intervals along the human genome (NCBI build 34), except for annotated gaps, heterochromatic, telomeric and centromeric regions. Y chromosomal SNPs were also excluded. Variation of the interval size between 5 and 10 kb for chromosomes 3 and 19 did not yield notably different results (data not shown). It may be argued that, in many instances, only markers located in gene-coding regions are of practical interest for genome-wide disease association studies. In order to take this issue into account, 'coding' regions were defined here as all sequences containing one of the 'RefSeq' genes of the Golden Path (http://genome.ucsc.edu), including exons, introns and 10 kb of flanking sequence. Marker sets were evaluated on the basis of publicly available genotype data (Table 1). Our analyses included CEPH samples from Northern and Western Europe (CEU), from Yoruba in Nigeria (YRI) and from Japanese and Han Chinese people (JPT + CHB).

Table 1. Sources of marker and genotype data

Marker set	SNPs	Used SNPs[a]	*N*	URL
HapMap (release 19)	3 719 872	2 496 932	60	www.hapmap.org
Affymetrix 100K	115 353	104 081	60	www.affymetrix.com
Affymetrix 500K	500 568	448 867	60	www.affymetrix.com
Illumina 100K	109 150	104 365[b]	32[c]	www.illumina.com
Illumina 300K	315 510	315 316[b]	60	www.illumina.com
Illumina 550K	548 944	527 207[b]	60	www.illumina.com

Column 'N' refers to the number of unrelated individuals (CEU, YRI. JPT + CHB) for whom genotypes were available. Founder individuals were used whenever possible. All data were retrieved from the listed URLs. Since HapMap release 19 was based upon the NCBI build 35 genome assembly, all marker positions were transformed accordingly.
[a]Excluding non-biallelic SNPs.
[b]Excluding SNPs that could not be identified in HapMap.
[c]Genotypes were available for only 32 of the Caucasian individuals.

Swept radii $1/\varepsilon$ were estimated for different genomic regions on the basis of the available HapMap genotype data. As is exemplified by chromosomes 12 and 19 in the CEU population (Figures 1A and 2A), the distribution of $1/\varepsilon$ was found to vary considerably along chromosomes and therefore resembled recently published recombination plots in this respect (2). The median $1/\varepsilon$ of ~500 kb corresponds closely to previous estimates (11). A graphical representation of all swept radii and τ values obtained in the present study is available at http://www.ikmb.unikiel.de/snpselection. In the following, our results will be exemplified by a more detailed consideration of chromosomes 12 and 19, which are typical in terms of their size and gene density.

When all 180 613 HapMap SNPs on chromosome 12 were included in the analysis, τ values larger than 0.5 were obtained for most of the chromosome (Figure 1C). By contrast, the 5253 chromosome 12 markers of the Affymetrix 100K set left many intervals with τ close to 0, indicating low efficacy (Figure 1B). Similar results were obtained for chromosome 19 (Figure 2). Figures 3 and 4 provide an overview of the distribution of τ along the coding' regions and the full genomic sequences of the two chromosomes. When all HapMap SNPs were included, the median τ values obtained were 0.70 (interquartile range: 0.56–0.82) for chromosome 12 and 0.66 (interquartile range: 0.52–0.78) for chromosome 19. By contrast, the best commercial marker sets yielded a median τ of 0.59 (interquartile range: 0.45–0.73) for chromosome 12, and of 0.56 (interquartile range: 0.41–0.70) for chromosome 19 in the case of Illumina 550K, and of 0.52 (interquartile range: 0.36–0.67) for chromosome 12 and of 0.41 (interquartile range: 0.26–0.58) for chromosome 19 with the Affymetrix 500K set.

A comparison of the two commercially available 100K sets revealed the impact of both, the genotyping

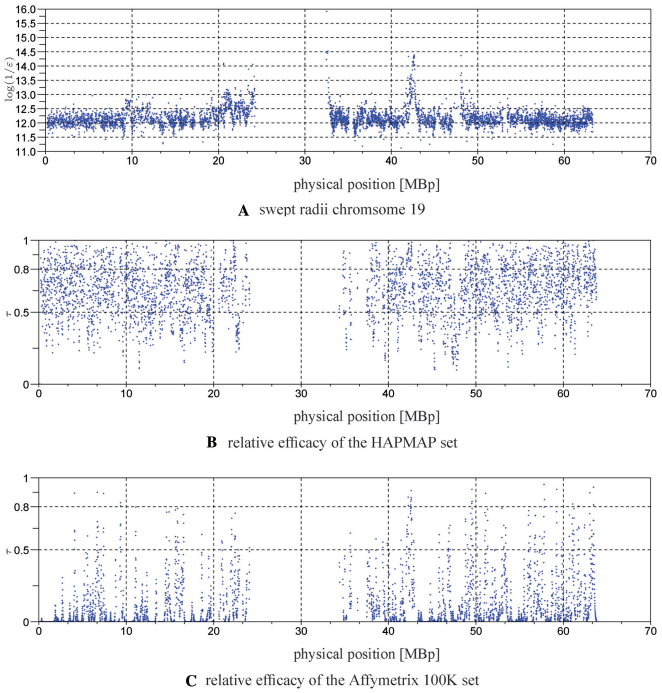

Figure 1. Distribution of local LD and SNP set efficacy on chromosome 19 in the CEU population. Panel **A**: swept radius $1/\varepsilon$ as estimated around each marker from the HapMap genotype data (median $1/\varepsilon$: 191 kb, interquartile range: 166–230 kb). Panel **B**: relative efficacy τ of the HapMap set, calculated at 10 kb intervals, excluding gaps, centromers, telomers and heterochromatin. Panel **C**: as Panel B, but for the Affymetrix 100K set. Note: physical positions (in Megabases, Mb) are given according to NCBI build 35.

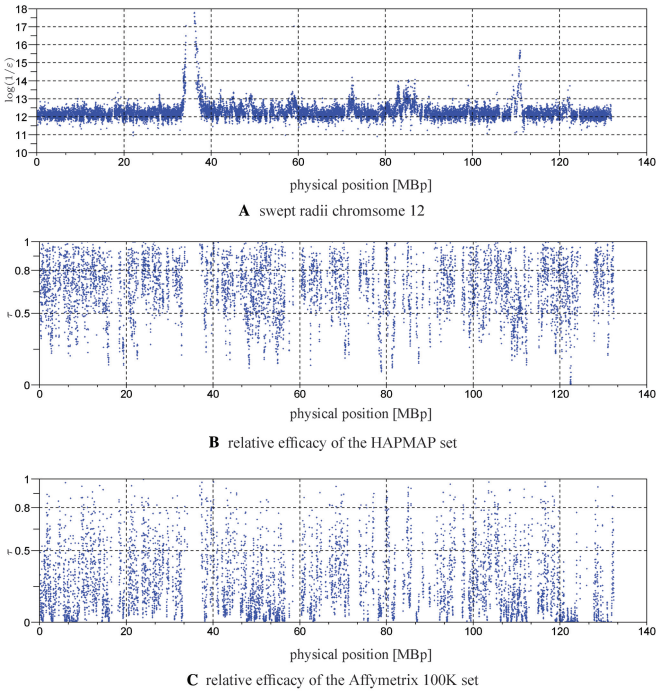

Figure 2. Distribution of local LD and SNP set efficacy on chromosome 12 in the CEU population. Panel **A**: swept radius $1/\varepsilon$ as estimated around each marker from the HapMap genotype data (median $1/\varepsilon$: 181 kb, interquartile range: 159–214 kb). Panels **B** and **C**: see legend to Figure 1.

technology and the selection strategy upon the mapping efficacy. If only the coding sequence was considered on chromosome 12, the median τ for Affymetrix 100K was 0.21 (interquartile range: 0.08–0.41), as compared to 0.44 (interquartile range: 0.27–0.61) for Illumina 100K (Figure 4). The Illumina 100K set, designed primarily for a good coverage of sequences containing annotated transcripts, provides essentially the same efficacy for the coding sequence on this gene-rich chromosome as the Affymetrix 500K set (median τ: 0.41, interquartile range: 0.26–0.58). Similar, albeit less pronounced results were obtained for chromosome 12 (Figure 3). A genome-wide overview of the efficacy of all SNP sets is given in Table 2 and, on a chromosome-wise basis, in Figure 5.

Let C_x denote the local coverage of a chromosome or chromosomal region at relative efficacy x, achieved by a particular marker set (i.e. C_x equals the proportion of a given genomic region for which $\tau \geq x$). For the coding regions of chromosome 12, for example, $C_{0.5} = 0.16$ for the Affymetrix 100K set and $C_{0.5} = 0.48$ for Affymetrix 500K (Figure 3). This means that the two sets cover 16 and 48% of the gene containing sequence, respectively, at 50% or higher relative efficacy. At 80% relative efficacy, the respective figures decrease to 2 and 8%, respectively. A genome-wide overview of the coverage of the different marker sets at 50 and 80% efficacy is given in Table 2 and, on a chromosome-wise basis, in Figures 6 and 7.

The HapMap markers provide the 'gold standard' for the currently achievable coverage of the human genome with informative SNPs. If a fully flexible genotyping technology were available, optimal SNP sets could thus be constructed from HapMap using, for example, entropy-based marker selection. As exemplified for chromosomes 12 (Figure 3) and 19 (Figure 4), such customized panels would significantly improve the coverage provided by a given number of markers. With 5253 SNPs on chromosome 12, which corresponds to the size of the respective Affymetrix 100K set, HapMap would yield $C_{0.5} = 0.81$, i.e. a more than four times higher coverage than the commercial product. Replacing the Affymetrix 500K set by a similarly sized HapMap set would increase $C_{0.5}$ from 0.48 to 0.81 whereas $C_{0.8}$ would increase from 0.07 to 0.21.

More detailed information about the present study can be found on our web server at http://www.ikmb.unikiel.de/snpselection. The same site also provides routines for the customized selection of optimal SNP sets from HapMap build 19, using the available Caucasian, Asian and Yoruba genotype data.

DISCUSSION

Justification of an entropy-based SNP selection framework

Currently available technologies do not allow full re-sequencing of the human genome in samples that are appropriately sized for mapping complex disease genes. Instead, the success of genome-wide association studies depends heavily upon the presence of sufficient LD

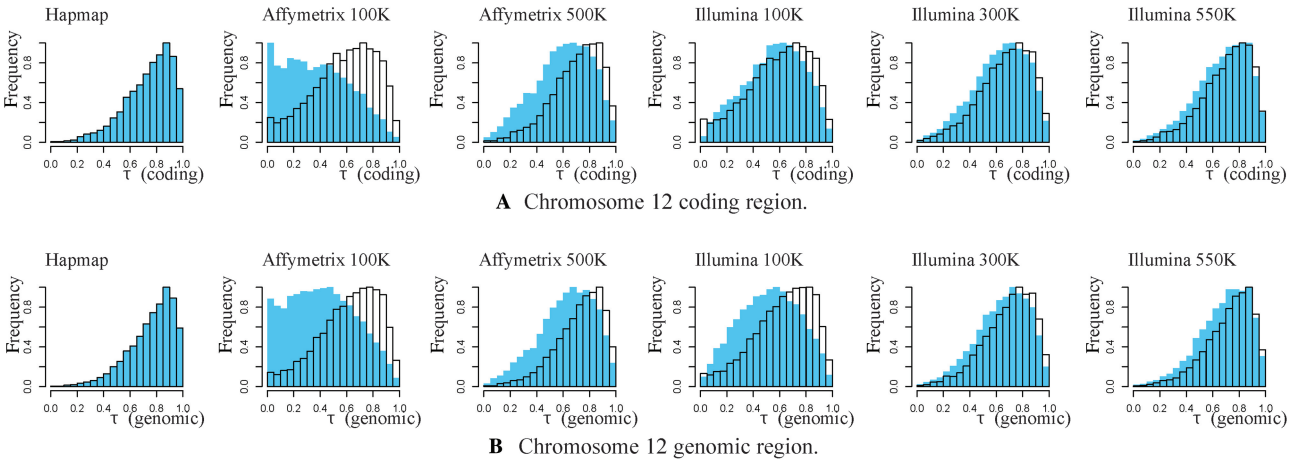

Figure 3. Relative efficacy of SNP sets on chromosome 12 in the CEU population. For each marker set, the blue histogram depicts the distribution of relative efficacy τ in the full genomic sequence and the coding regions, respectively (for definition, see main text). Frequencies have been normalized such that the modal frequency equals unity. The distribution of τ as obtained for a similarly sized, hypothetical marker set, constructed from HapMap by entropy-based marker selection, is included for each marker set (open histograms).

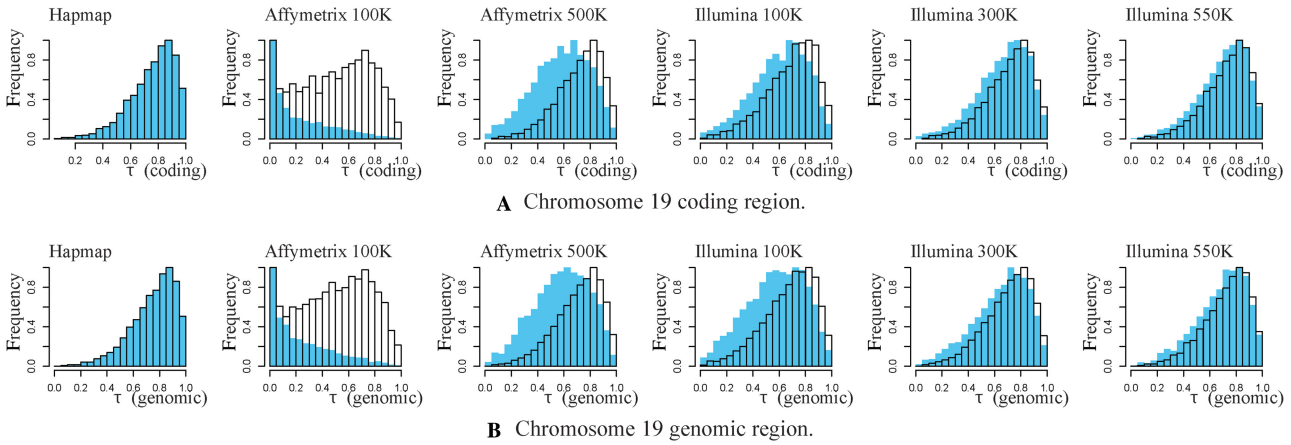

Figure 4. Relative efficacy of SNP sets on chromosome 19 in the CEU population. For details, see legend to Figure 3.

between the causal variant(s) and at least one marker in the study panel. Whilst the level of inter-marker LD may indeed be fully known, however, LD is inherently unknown in relation to the causal variant itself, and therefore has to be extrapolated. This implies that the markers of an ideal study panel should be selected in such a way as to maximize the information extracted about any possible location of a disease variant in the genome.

Under a model of spatially homogenous LD, with constant recombination and mutation rates and a common evolutionary history shared by all chromosomal regions, disease association markers would ideally be spread evenly along the genome. However, the systematic evaluation of both LD and local recombination rates has revealed an inherent non-uniformity of these characteristics (2,13,14). Thus, recombination rates differ between chromosomal segments and between populations, which implies that even closely linked genomic regions may be of substantially different ancestry in individuals from one and the same population (15). Consequently, the relationship between LD and physical distance is complex, and combinations of unevenly spaced SNPs may prove more informative than equally spaced markers, depending upon the genomic region of interest.(16)

Previous studies have suggested the existence of 'haplotype blocks', i.e. clearly identifiable chromosomal segments that are characterized by a reduced rate of recombination, low haplotype diversity and a high level of internal LD (2–4). In addition, haplotype-tagging SNPs (htSNPs) have been proposed to be capable of identifying haplotypes for substantially larger marker sets from within these blocks (17–19). The practical relevance of this block concept arises from the expectation that htSNPs extract sufficient information from an LD block with

Table 2. Median estimated efficacy and coverage of the human genome (excluding the Y chromosome) in different populations, provided by different marker sets.

Population	Marker set	Full genomic sequence			'Coding' regions		
		τ (interquartile range)	$C_{0.5}$	$C_{0.8}$	τ (interquartile range)	$C_{0.5}$	$C_{0.8}$
CEU	HapMap (total)	0.71 (0.57–0.84)	83%	33%	0.7 (0.55–0.83)	82%	30%
	Affymetrix 100K	0.26 (0.11–0.47)	23%	7%	0.23 (0.09–0.44)	20%	5%
	Affymetrix 500K	0.52 (0.35–0.69)	53%	13%	0.5 (0.34–0.67)	50%	12%
	Illumina 100K	0.37 (0.21–0.58)	33%	9%	0.44 (0.27–0.63)	42%	10%
	Illumina 300K	0.55 (0.4–0.71)	59%	14%	0.54 (0.39–0.7)	58%	13%
	Illumina 550K	0.6 (0.45–0.75)	68%	18%	0.6 (0.44–0.74)	66%	17%
YRI	HapMap (total)	0.74 (0.62–0.85)	91%	36%	0.73 (0.61–0.84)	90%	34%
	Affymetrix 100K	0.25 (0.1–0.45)	21%	6%	0.22 (0.08–0.42)	19%	5%
	Affymetrix 500K	0.53 (0.38–0.69)	55%	13%	0.51 (0.37–0.68)	53%	12%
	Illumina 100K	0.29 (0.15–0.49)	24%	7%	0.37 (0.22–0.56)	32%	7%
	Illumina 300K	0.52 (0.37–0.68)	53%	12%	0.51 (0.36–0.67)	52%	11%
	Illumina 550K	0.59 (0.45–0.74)	67%	16%	0.58 (0.44–0.73)	65%	15%
JPT + CHB	HapMap (total)	0.68 (0.52–0.82)	77%	28%	0.67 (0.51–0.81)	76%	26%
	Affymetrix 100K	0.22 (0.08–0.42)	19%	6%	0.19 (0.06–0.39)	17%	5%
	Affymetrix 500K	0.48 (0.31–0.66)	47%	12%	0.47 (0.3–0.65)	45%	11%
	Illumina 100K	0.27 (0.12–0.47)	23%	7%	0.34 (0.18–0.54)	29%	7%
	Illumina 300K	0.5 (0.33–0.67)	50%	12%	0.49 (0.33–0.66)	48%	11%
	Illumina 550K	0.56 (0.39–0.72)	60%	15%	0.55 (0.38–0.71)	58%	14%
CEU	Hyp. Affymetrix 100K	0.50 (0.32–0.68)	50%	12%	0.47 (0.29–0.65)	46%	10%
	Hyp. Affymetrix 500K	0.64 (0.49–0.78)	73%	22%	0.62 (0.47–0.76)	71%	19%
	Hyp. Illumina 100K	0.46 (0.25–0.63)	44%	7%	0.44 (0.24–0.62)	42%	7%
	Hyp. Illumina 300K	0.62 (0.46–0.76)	70%	19%	0.60 (0.44–0.74)	67%	17%
	Hyp. Illumina 550K	0.66 (0.52–0.79)	77%	24%	0.65 (0.50–0.78)	75%	21%
YRI	Hyp. Affymetrix 100K	0.55 (0.36–0.71)	57%	14%	0.52 (0.33–0.69)	53%	12%
	Hyp. Affymetrix 500K	0.69 (0.57–0.8)	85%	26%	0.68 (0.55–0.79)	82%	23%
	Hyp. Illumina 100K	0.51 (0.32–0.69)	52%	13%	0.49 (0.3–0.67)	48%	11%
	Hyp. Illumina 300K	0.66 (0.53–0.78)	79%	22%	0.64 (0.5–0.77)	75%	19%
	Hyp. Illumina 550K	0.69 (0.57–0.8)	85%	26%	0.68 (0.55–0.79)	83%	23%
JPT + CHB	Hyp. Affymetrix 100K	0.42 (0.2–0.63)	40%	10%	0.4 (0.19–0.6)	37%	9%
	Hyp. Affymetrix 500K	0.6 (0.42–0.76)	65%	19%	0.59 (0.4–0.74)	63%	17%
	Hyp. Illumina 100K	0.4 (0.18–0.61)	37%	10%	0.38 (0.16–0.58)	35%	8%
	Hyp. Illumina 300K	0.53 (0.28–0.71)	54%	15%	0.51 (0.27–0.69)	52%	13%
	Hyp. Illumina 550K	0.59 (0.4–0.74)	63%	15%	0.58 (0.39–0.73)	62%	14%

All estimates refer to NCBI build 35, excluding annotated gaps, centromers, heterochromatin and telomers. 'Coding' regions were defined by the 'RefSeqs' provided in the Golden Path (http://genome.ucsc.edu), including introns, exons and 10 kb of flanking sequence.
τ genome-wide median of the relative efficacy; $C_{0.5}$, ($C_{0.8}$): percentage of the autosomal genome covered with $\tau \geq 0.5$ ($\tau \geq 0.8$); Hyp.: hypothetical.

respect to co-ancestry while, at the same time, reducing genotyping costs (2–4). A number of computational methods for the construction of htSNP sets have been developed (16,19,20) but for these techniques to be efficient, detailed knowledge of the extended haplotype frequency distribution in the population of interest is required. Moreover, the size and location of haplotype blocks depend critically upon the SNP density and the method of marker selection (13,14,21–24). Therefore, haplotype tagging appears feasible only when large samples and appropriate family structures are available for the necessary (deterministic or probabilistic) haplotype assignments, the reliability of which decreases with the number and complexity of the haplotypes present (25–27).

The idealized picture of static LD blocks, separated by hot spots of recombination, (28) has recently been challenged by new insights into the biology of meiotic recombination (2–4). The correlation between physical and genetic distance is weak below the centi-Morgan level so that the inference of marker genotypes from htSNP haplotypes is far from being reliable (24). Moreover, block-like structures may even occur merely because of genetic drift (29). It thus appears as if the tacit assumption underlying the use of the haplotype block concept for disease association mapping, namely that all genetic variation in a block follows the same hierarchical pattern, is often not fulfilled. As a consequence, the usefulness of htSNPs for such studies has generally been questioned (30–32).

SNP selection based upon pairwise LD alone has been suggested to avoid the conceptual and computational problems of extended haplotype (or 'block') approaches. The use of some SNPs as proxies for other SNPs that are in high LD with the former (2–4), measured by r^2, reduces the redundancy of a SNP set. Thresholds for r^2 of at least 0.8 are generally regarded as sufficient to provide good marker coverage for association studies (21,33–38). The rationale underlying the pairwise approach is the expectation that high inter-marker LD translates into high LD between some of the markers and potentially causative variants, an assumption that is however unlikely to hold

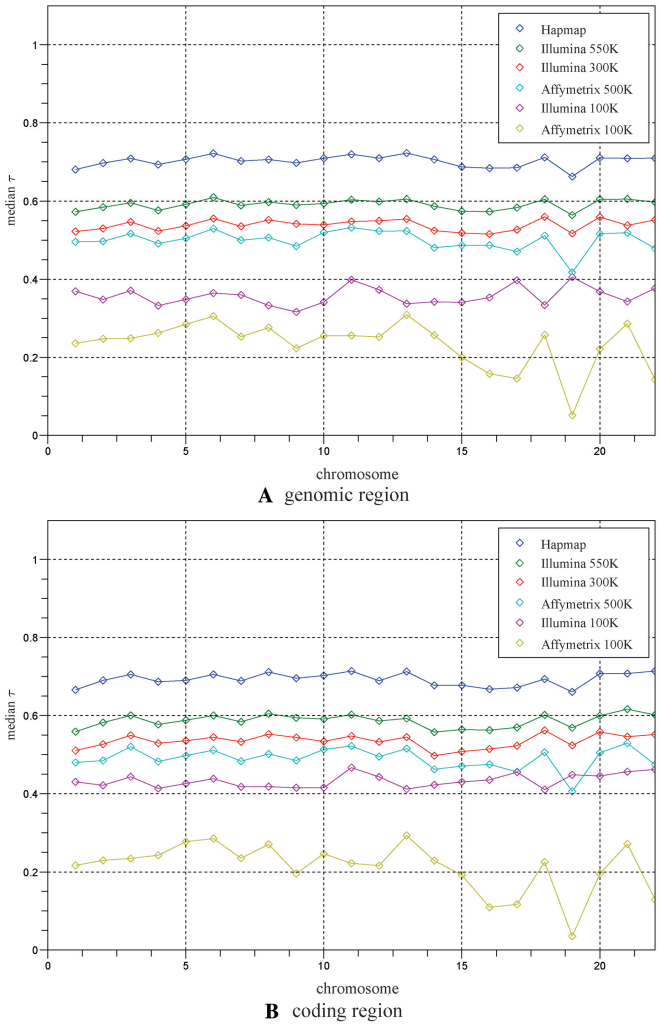

Figure 5. Chromosome-specific estimates of relative SNP set efficacy in full genomic (Panel **A**) and coding (Panel **B**) sequences. Chromosome-wide median τ values and interquartile ranges obtained for the CEU population are plotted in chromosomal order.

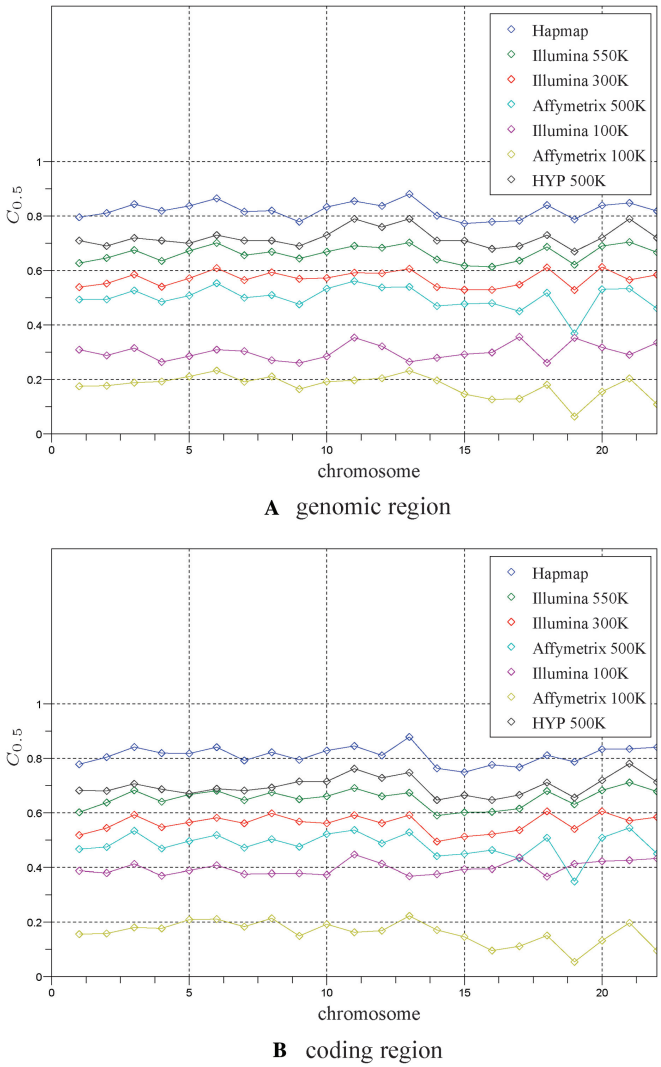

Figure 6. SNP set coverage of full genomic (Panel **A**) and coding (Panel **B**) sequences at 50% relative efficacy. The chromosome-wide coverage $C_{0.5}$ is plotted in chromosomal order. HYP 500K: hypothetical, optimal marker set constructed from HapMap so as to include the same number of SNPs per chromosome as the Affymetrix 500K set.

true in general (2–4). Selection of SNPs based upon pairwise LD alone is therefore likely to perform well only with a particularly high and uniform SNP density (6). Irrespective of the approach taken, the inherently unknown LD between markers and unknown causal variants has to be extrapolated in one way or another from both physical distance and the local strength of LD. However, marker selection based upon pairwise LD alone does not take distance or individual marker informativity into account. As a consequence, simple pairwise 'haplotype tagging' potentially leads to inhomogeneous marker spacing with less than maximum efficacy.

Here, we have adapted a recently proposed method for selecting maximally informative marker sets for association studies (6) to a genome-wide comparison of marker sets. The original approach combines the information content, physical spacing and pairwise LD of individual markers with information on the local LD structure, extracted from available data in the form of swept radii (10,11). All of these determinants are included in a single, position-specific utility measure that corresponds to the distance-weighted haplotype entropy of the marker set, approximated however by a pairwise score of the same form (see Methods section). The approach is therefore not affected by the computational and conceptual problems of block-based methods and, at the same time, takes physical distance and local LD structure into account when extrapolating LD between markers and causal variants from pairwise inter-marker LD. An extension of the approach has led to the development of a quantitative criterion (τ) that approximates the efficacy of a given marker set to map a disease-causing variant at a position of interest. It should be emphasized that the interpretation of τ as a measure of efficacy is only valid in relative terms, i.e. by comparison to the inclusion of a maximally

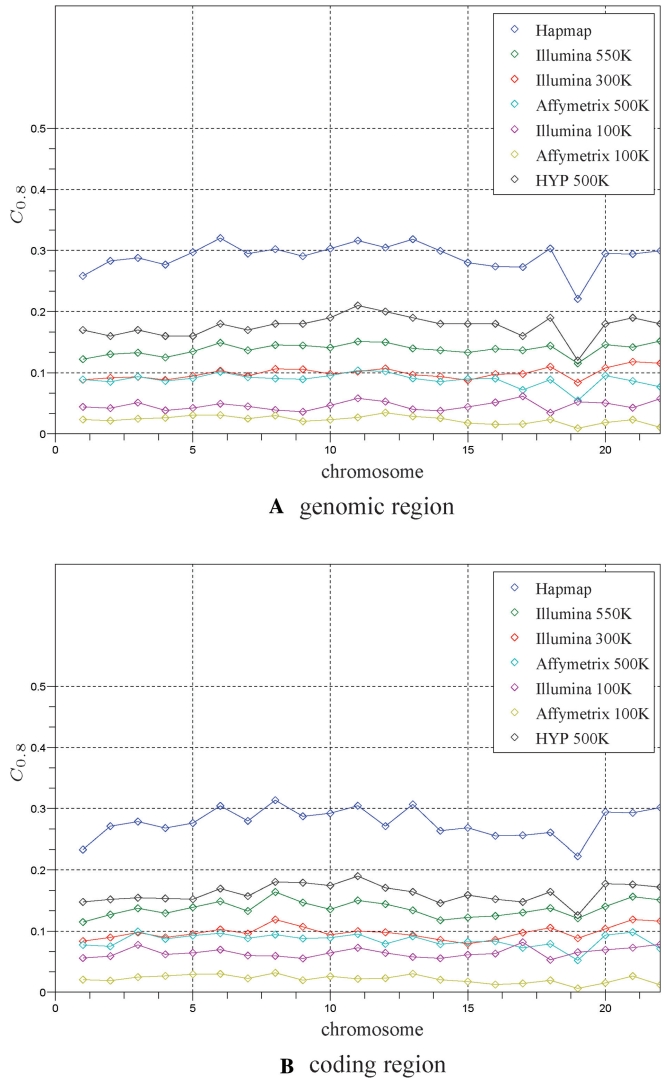

Figure 7. SNP set coverage of full genomic (Panel **A**) and coding (Panel **B**) sequences at 80% relative efficacy. The chromosome-wide coverage $C_{0.8}$ is plotted in chromosomal order. HYP 500K: hypothetical, optimal marker set constructed from HapMap so as to include the same number of SNPs per chromosome as the Affymetrix 500K set.

polymorphic SNP at the site of the causal variant. In general, since the properties of the underlying disease model are unknown, no marker-based quantity can on its own provide information about the absolute power of a marker set to map genetic variants underlying a given phenotype.

Quality of currently available marker sets

Owing to recent successes (39) and its theoretical appeal (1), significant funds have been allocated to the concept of genome-wide association analysis in the context of various phenotypes. Researchers are however facing the practical problem of choosing the 'right' genotyping technology. In many countries, universal control genotype pools are in the process of being established, and these pools will pre-determine the choice of technology for future studies. Of the currently available marker sets, the Affymetrix 500K ($C_{0.5} = 0.68$, $C_{0.8} = 0.19$) and Illumina 550K ($C_{0.5} = 0.79$, $C_{0.8} = 0.29$) products provide the best genomic coverage in Caucasians. The Illumina 550K marker set provides a higher coverage than the 500K Affymetrix set, probably because of the higher flexibility of the Illumina genotyping technology. Pronounced differences between full genomic and 'coding' region coverage were only observed for the 100K sets, probably because of the relatively small marker numbers. The good 'coding' region coverage provided by the Illumina 100K set highlights the fact that this panel was primarily designed for gene-based association mapping. It should be emphasized, however, that all of the above conclusions were based upon the assumption that all markers were callable, and that practical factors such as genotyping quality, departure from Hardy–Weinberg equilibrium and DNA requirements could be neglected. Furthermore, interesting differences became apparent in terms of in different ethnic groups. Whilst their relative efficacy was approximately the same in the Caucasian and African populations, SNP coverage was notably poorer for all products for the East Asian populations.

The analytical method used here to compare the utility of different marker sets provides a means to weight the costs and benefits of closing gaps in a given marker set. Additional genotyping costs incurred by a flexible (and thus more expensive) genotyping method can be contrasted directly with the relative efficacy gained from using additional, customized SNPs. If genotyping costs would be negligible, the complete current HapMap set would provide 90% coverage of the genome with at least 50% relative efficacy, and 47% coverage with at least 80% relative efficacy. These figures represent the gold standard with which all other marker panels have to be compared. Interestingly, when our entropy-based SNP selection approach was used to construct an optimum SNP set, the size of the Affymetrix 500K product from HapMap, this technology-independent, hypothetical set would nearly double the coverage at 80% relative efficacy.

In summary, we have devised a methodology that helps researchers make rational choices between different marker sets for genome-wide disease association studies and to assess the trade-off between genotyping costs and gain in power when expanding existing marker sets. Furthermore, use of the τ criterion facilitates judging the position-specific 'completeness' of a genome-wide association study and may thus help to improve the practicability of complex disease gene mapping.

ACKNOWLEDGEMENTS

This study was supported by the German Federal Ministry of Education and Research as part of the National Genome Research Network (01GS02105, 0313437A) and the MediGrid project (01AK803G), and by the German Research Council (Ha 3091/1-2). We are most grateful to Ulf Leser, Humboldt-University,

Berlin, for helpful discussions and to Uwe Mordhorst and Marcus Will, Christian-Albrechts-University, Kiel, for computing support. Funding to pay the Open Access publication charges for this article was provided by the Medical Faculty of the University of Kiel.

Conflict of interest statement. None declared.

REFERENCES

1. Risch,N. and Merikangas,K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
2. Myers,S., Bottolo,L., Freeman,C., McVean,G. and Donnelly,P. (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science*, **310**, 321–324.
3. Jeffreys,A.J. and Neumann,R. (2005) Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Hum. Mol. Genet.*, **14**, 2277–2287.
4. Jeffreys,A.J., Neumann,R., Panayi,M., Myers,S. and Donnelly,P. (2005) Human recombination hot spots hidden in regions of strong marker association. *Nat. Genet.*, **37**, 601–606.
5. Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech.l J.*, **27**, 379–423.
6. Hampe,J., Schreiber,S. and Krawczak,M. (2003) Entropy-based SNP selection for genetic association studies. *Hum. Genet.*, **114**, 36–43.
7. Altshuler,D., Brooks,L.D., Chakravarti,A., Collins,F.S., Daly,M.J. and Donnelly,P. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
8. Matsuzaki,H., Dong,S., Loi,H., Di,X., Liu,G., Hubbell,E., Law,J., Berntsen,T., Chadha,M. *et al.* (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods*, **1**, 109–111.
9. Murray,S.S., Oliphant,A., Shen,R., McBride,C., Steeke,R.J., Shannon,S.G., Rubano,T., Kermani,B.G., Fan,J.B. *et al.* (2004) A highly informative SNP linkage panel for human genetic studies. *Nat. Methods*, **1**, 113–117.
10. Morton,N.E. and Collins,A. (1998) Tests and estimates of allelic association in complex inheritance. *Proc. Natl Acad. Sci. USA*, **95**, 11389–11393.
11. Morton,N.E., Zhang,W., Taillon-Miller,P., Ennis,S., Kwok,P.Y. and Collins,A. (2001) The optimal measure of allelic association. *Proc. Natl Acad. Sci. USA*, **98**, 5217–5221.
12. Collins,A., Lonjou,C. and Morton,N.E. (1999) Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA*, **96**, 15173–15177.
13. Gabriel,S.B., Schaffner,S.F., Nguyen,H., Moore,J.M., Roy,J., Blumenstiel,B., Higgins,J., DeFelice,M., Lochner,A. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.
14. Sawyer,S.L., Mukherjee,N., Pakstis,A.J., Feuk,L., Kidd,J.R., Brookes,A.J. and Kidd,K.K. (2005) Linkage disequilibrium patterns vary substantially among populations. *Eur. J. Hum. Genet.*, **13**, 677–686.
15. Rosenberg,N.A., Pritchard,J.K., Weber,J.L., Cann,H.M., Kidd,K.K., Zhivotovsky,L.A. and Feldman,M.W. (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.
16. Johnson,G.C., Esposito,L., Barratt,B.J., Smith,A.N., Heward,J., Di Genova,G., Ueda,H., Cordell,H.J., Eaves,I.A. *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.*, **29**, 233–237.
17. Patil,N., Berno,A.J., Hinds,D.A., Barrett,W.A., Doshi,J.M., Hacker,C.R., Kautzer,C.R., Lee,D.H., Marjoribanks,C. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.
18. Zhang,K., Calabrese,P., Nordborg,M. and Sun,F. (2002) Haplotype block structure and its applications to association studies: power and study designs. *Am. J. Hum. Genet.*, **71**, 1386–1394.
19. Zhang,K., Deng,M., Chen,T., Waterman,M.S. and Sun,F. (2002) A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl Acad. Sci. USA*, **99**, 7335–7339.
20. Halperin,E., Kimmel,G. and Shamir,R. (2005) Tag SNP selection in genotype data for maximizing SNP prediction accuracy. *Bioinformatics*, **21**(Suppl. 1), i195–i203.
21. Ke,X., Durrant,C., Morris,A.P., Hunt,S., Bentley,D.R., Deloukas,P. and Cardon,L.R. (2004) Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Hum. Mol. Genet.*, **13**, 2557–2565.
22. Kidd,J.R., Pakstis,A.J., Zhao,H., Lu,R.B., Okonofua,F.E., Odunsi,A., Grigorenko,E., Tamir,B.B., Friedlaender,J. *et al.* (2000) Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. *Am. J. Hum. Genet.*, **66**, 1882–1899.
23. Sun,X., Stephens,J.C. and Zhao,H. (2004) The impact of sample size and marker selection on the study of haplotype structures. *Hum. Genomics*, **1**, 179–193.
24. Nothnagel,M. and Rohde,K. (2005) The effect of SNP marker selection on patterns of haplotype blocks and haplotype frequency estimates. *Am. J. Hum. Genet.*, **77**, 988–998.
25. Becker,T. and Knapp,M. (2002) Efficiency of haplotype frequency estimation when nuclear familiy information is included. *Hum. Hered.*, **54**, 45–53.
26. Douglas,J.A., Boehnke,M., Gillanders,E., Trent,J.M. and Gruber,S.B. (2001) Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nat. Genet.*, **28**, 361–364.
27. Schaid,D.J. (2002) Relative efficiency of ambiguous vs. directly measured haplotype frequencies. *Genet. Epidemiol.*, **23**, 426–443.
28. Goldstein,D.B. (2001) Islands of linkage disequilibrium. *Nat. Genet.*, **29**, 109–111.
29. Wang,N., Akey,J.M., Zhang,K., Chakraborty,R. and Jin,L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.*, **71**, 1227–1234.
30. Crawford,D.C., Carlson,C.S., Rieder,M.J., Carrington,D.P., Yi,Q., Smith,J.D., Eberle,M.A., Kruglyak,L. and Nickerson,D.A. (2004) Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am. J. Hum. Genet.*, **74**, 610–622.
31. Zhai,W., Todd,M.J. and Nielsen,R. (2004) Is haplotype block identification useful for association mapping studies? *Genet. Epidemiol.*, **27**, 80–83.
32. Pritchard,J.K. and Cox,N.J. (2002) The allelic architecture of human disease genes: common disease-common variant . . . or not? *Hum. Mol. Genet.*, **11**, 2417–2423.
33. Wang,W.Y., Barratt,B.J., Clayton,D.G. and Todd,J.A. (2005) Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.*, **6**, 109–118.
34. Carlson,C.S., Eberle,M.A., Rieder,M.J., Yi,Q., Kruglyak,L. and Nickerson,D.A. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, **74**, 106–120.
35. Carlson,C.S., Eberle,M.A., Rieder,M.J., Smith,J.D., Kruglyak,L. and Nickerson,D.A. (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat. Genet.*, **33**, 518–521.
36. Lowe,C.E., Cooper,J.D., Chapman,J.M., Barratt,B.J., Twells,R.C., Green,E.A., Savage,D.A., Guja,C., Ionescu-Tirgoviste,C. *et al.* (2004) Cost-effective analysis of candidate genes using htSNPs: a staged approach. *Genes. Immun.*, **5**, 301–305.
37. Chapman,J.M., Cooper,J.D., Todd,J.A. and Clayton,D.G. (2003) 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.
38. Wang,W.Y. and Todd,J.A. (2003) The usefulness of different density SNP maps for disease association studies of common variants. *Hum. Mol. Genet.*, **12**, 3145–3149.
39. Klein,R.J., Zeiss,C., Chew,E.Y., Tsai,J.Y., Sackler,R.S., Haynes,C., Henning,A.K., Sangiovanni,J.P., Mane,S.M. *et al.* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385–389.
40. Fleiss,J.L. (2003) *Statistical Methods for Rates and Proportions, 3rd edn.* John Wiley & Sons, New York, USA.

APPENDIX

In general, the sample size n required to detect the difference between proportions π_1 and π_2 by means of a χ^2 test can be approximated by

$$n = \frac{[z_{1-\alpha/2} \cdot \sqrt{2\pi(1-\pi)} + z_{1-\beta} \cdot \sqrt{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}]^2}{(\pi_1 - \pi_2)^2} \tag{A.1}$$

where $\pi = (\pi_1 + \pi_2)/2$, α and $1-\beta$ are the significance level and power of the applied test, respectively, and $z_{1-\alpha/2}$ and $z_{1-\beta}$ denote the corresponding quantiles of the Gaussian distribution (40). If q is the minor allele frequency of marker X, and if the two alleles of marker Z are equally frequent, then the corresponding haplotype frequency matrix equals

$$\mathrm{P} = \left(p_{ij}\right)_{i,j=1..2} = \begin{bmatrix} \frac{1}{2}\pi_1 & \frac{1}{2}(1-\pi_1) \\ \frac{1}{2}\pi_2 & \frac{1}{2}(1-\pi_2) \end{bmatrix} \tag{A.2}$$

with

$$\frac{1}{2}\pi_1 + \frac{1}{2}\pi_2 = 1 - q. \tag{A.3}$$

Furthermore, since $Q = p_{11} + p_{12} = 0.5\pi_1 + 0.5(1-\pi_1) = 0.5$ and $R = p_{11} + p_{12} = 0.5\pi_1 + 0.5\pi_2$, it follows that

$$\rho = \frac{\det|\mathrm{P}|}{Q \cdot (1-R)} = \frac{\frac{1}{2}\pi_1 \cdot \frac{1}{2}(1-\pi_2) - \frac{1}{2}\pi_2 \cdot \frac{1}{2}(1-\pi_1)}{\frac{1}{2} \cdot \left(1 - \frac{1}{2}\pi_1 - \frac{1}{2}\pi_2\right)} = \frac{\pi_1 - \pi_2}{2 - \pi_1 - \pi_2}. \tag{A.4}$$

Solving Equations (A.3) and (A.4) for π_1 and π_2 yields $\pi_1 = 1 - q(1-\rho)$ and $\pi_2 = 1 - q(1+\rho)$, so that $\pi = 1 - q$ and $\pi_1 - \pi_2 = 2q\rho$. Replacing π_1, π_2 and π by these expressions in formula (A.1) yields

$$n = \frac{[z_{1-\alpha/2} \cdot \sqrt{2q(1-q)} + z_{1-\beta} \cdot \sqrt{2q(1-q[1+\rho^2])}]^2}{(2q\rho)^2} \sim \frac{[z_{1-\alpha/2} + z_{1-\beta}]^2}{2} \cdot \frac{1-q}{q\rho^2} \tag{A.5}$$

for sufficiently small ρ. This proves formula (10) of the main text.