The usefulness of different density SNP maps for disease association studies of common variants

William Y. S. Wang and John A. Todd*

JDRF/WT Diabetes and Inflammation Laboratory, Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 2XY, UK

Received August 5, 2003; Revised and Accepted September 25, 2003

Large-scale discovery and validation of single-nucleotide polymorphisms (SNPs) facilitates indirect association mapping. It has recently been estimated that, in Europeans, 77% of all SNPs with frequency of 10% or more could be ascertained through linkage disequilibrium (LD) by genotyping variants in the database dbSNP. Using a sampling approach from 73 genes with near complete SNP maps, we show here the usefulness of SNP maps at different densities and the large variability of SNP coverage in different genomic regions. While even sparse SNP maps are of some value to genetic mapping, in order to undertake disease association studies providing at least 80% of SNPs in 90% of genes, much denser maps need to be constructed, at more than one SNP per kb in some regions.

INTRODUCTION

Indirect genetic association mapping aims to detect causative disease variants via their non-random association, LD, with genotyped SNPs (1,2). This approach is currently targeted at common SNPs, as they make at least some contribution to human diseases (3), and allow statistically powered studies in contrast to rarer variants. It is unclear, however, what proportion of the estimated six million common (frequency \(>10\%\)) European SNPs in the human genome (4,5) need to be first discovered and validated to ensure highly comprehensive large-scale association mapping. The construction of SNP maps allows LD analyses and selection of an economic and statistically powerful set of haplotype tag SNPs (htSNPs) for genotyping in association studies (6,7).

Until recently, estimates on the density requirements have been based on relatively sparse sampling of the SNP content of the genome (8–10). Carlson et al. (5) provided SNP data from 50 completely resequenced genes allowing a more accurate assessment of SNP ascertainment. They estimated that from the 2.7 million unique SNPs then available in dbSNP (11), approximately 1.35 million were common in Europeans, and genotyping of these common SNPs would ascertain 77% of all common European SNPs at \(r^2 \geq 0.8\). Since the number of disease-associated variants in the genome is likely to be large, such rates of SNP ascertainment would be of significant value to genetic researchers. However, owing to variations in LD and in SNP sampling, incomplete maps will inevitably contain regions with very poor SNP capture, and this has not been previously investigated. Accounting for LD and sampling variations, we here estimate the SNP coverage in genes for different map densities via simulated sampling from the near complete SNP maps of the UW-FHCRC Variation Discovery Resource database (http://pga.mbt.washington.edu/).

RESULTS

We selected from the database 73 autosomal gene segments that were greater than 10 kb in length and at least 90% resequenced, yielding 2523 common European SNPs and 1586 kb in genomic length (Supplementary Material, Table 1). SNP genotypes were of 23 European American individuals.

From the set of all SNPs, we created subsets of mapped (observed) SNPs by sampling based on density. To understand how different sampling approaches may affect the final outcome, we devised three sampling strategies: (i) random sampling from the pool of all genes; (ii) random sampling within each gene to a given density; and (iii) sampling with the aim of even spacing by choosing each SNP to be as close to a pre-set evenly spaced position as possible.

*To whom correspondence should be addressed. Tel: +44 1223762101; Fax: +44 1223762102; Email: john.todd@cimr.cam.ac.uk
SNP coverage

Figure 1 shows the proportion of overall SNPs ascertained directly or indirectly as functions of target SNP densities. Sampling one common SNP (frequency ≥10%) per 5 kb (0.2 SNPs per kb in the figures) would capture more than 50% of SNPs at $r^2 ≥ 0.8$, and greater than 80% of SNPs in 22 genes (30%) in an evenly spaced map (Fig. 2A). Sampling one SNP per 2.5 kb ascertained 76% of SNPs, concurring with the previous estimate for the SNPs in dbSNP (5), and provided more than 80% SNP coverage for 38 of the 73 genes (52%; Fig. 2B). However, owing to variation in LD between different genomic regions, to achieve comprehensive disease mapping for common variants is more difficult. In the evenly spaced sampling of one SNP per 2.5 kb, seven genes (10%) were less than 50% covered. To obtain 80% coverage of SNPs in 80% of genes, sampling at densities of up to one SNP per 1.5 kb would be required in some genomic regions (Fig. 2C). Random sampling yielded lower levels of SNP capture in genes than the evenly spaced method.

Number of htSNPs

We next examined the number of htSNPs in relation to SNP map density. For each gene, starting from all mapped common SNPs, we minimized the SNPs to retain only one variant from each pair with $r^2 ≥ 0.8$. The resultant htSNPs should have similar rates of ascertainment of unobserved SNPs to the set of mapped SNPs. Using evenly spaced SNP maps, 366 and 483 htSNPs were chosen for initial SNP map densities of one SNP per 2.5 kb and one SNP per 1.5 kb, respectively (43 and 53% reduced genotyping). If these results are extended to the entire genome, then 760 000 and 1.0 million SNPs would be assayed for association studies based on the respective SNP maps. For the complete SNP map based on these 73 genes, 695 htSNPs were chosen, equating to 72% reduced genotyping, translating to 1.4 million htSNPs genome-wide.

DISCUSSION

We have estimated the usefulness of different density SNP maps and shown the variability between genomic regions. The varied SNP density requirements in different parts of the genome is consistent with previous observations of LD and SNP density variations (10,12–14). For genome-wide SNP map construction by ascertaining SNPs randomly from a relatively small number of individuals, sampling errors will also be important. The uncertainty regarding the actual number of SNPs and haplotypes in any particular region will make hierarchical SNP sampling strategies challenging. Nevertheless, the latest build 116 of dbSNP has 5.9 million
unique SNPs and will be of a better resource to begin SNP map
collection than the 2.7 million SNPs previously available (5,11).
We suggest that, even in Europeans, many more than
2.7 million SNPs will need to be genotyped in the mapping
panel to achieve good ascertainment for a large proportion of
genes.

A number of factors could lead to differences in SNP map
density estimates. Indirect testing of unobserved variants using
haplotypes, and not simply via pair-wise $r^2$ as we have done
here, should lower the map density requirements. However,
increased potential search space and degrees of freedom in
haplotype based tests may affect statistical power for detecting
disease association (15). The study of lower frequency SNPs
and of African populations will also require denser maps
(5,10). In contrast, regions of strong LD with limited haplotype
diversity would provide better SNP coverage for the same
density and lower htSNP requirements. Simulations using a
coalessent model showed that studying 100 kb genomic seg-
ments provides better overall coverage than those of our size,
but by less than 3% for SNP densities of greater than one SNP
per 5 kb (data not shown). This difference was halved with the
use of only 23 unphased individuals in the smaller segments
owing to upward biases in observed $r^2$ from small number of
individuals (see Supplementary Material Figures).

Our estimates of the number htSNPs confirmed that
significant genotyping reductions in disease association studies
can be achieved. Furthermore, the numbers we provided are
likely to be the upper limits. The use of haplotypes for htSNP
selection should be more efficient than using only pair-wise $r^2$,
but requires larger sample sizes for reliability (7). Data from
our own laboratory (in Supplementary Material, Table 2)
suggest that haplotype methods (15) may lead to ~30% further
reduction in genotyping. Larger regions of strong LD may also
be important. Therefore, the actual number of htSNPs required
may be less than what we suggested, perhaps a number
somewhat less than one million.

Given the likely large number of genetic variants for common
diseases, even SNP maps of one common SNP per 2.5 kb

Figure 2. Variation of SNP capture between genes for different marker densities. Marker densities of one SNP per 5 (A), 2.5 (B), 1.5 (C) and 1 kb (D) are shown.
For each sampling methods, random overall (solid bars), random within genes (patterned bars) or evenly spaced (open bars), the 73 genes studied were divided into
10 bins based on the rates of SNP ascertainment at the marker density.
covering a moderate proportion of common SNPs would be of value to begin association mapping (Fig. 2B). In the long term, however, geneticists searching chromosome regions to detect disease associations need to be able to confidently determine the involvement of at least the common SNPs, owing to uncertainties about the number, penetration and allelic frequencies of causal variants (16). We have shown that to achieve comprehensive and robust SNP maps, density focused, PCR-based, SNP discovery efforts at the minimum of one common SNP/kb density are required for parts of the genome. It will, therefore, be beneficial for disease mapping studies, both direct and indirect, that the on-going large-scale SNP map effort be integrated with the parallel construction of a SNP map of the 200 000–300 000 exons of the human genome by comprehensive resequencing.

**MATERIALS AND METHODS**

**Data set**

SNP data was derived from the UW-FHCRC Variation Discovery Resource database as of 1 June 2003. Only autosomal gene segments that were greater than 10 kb in length and more than 90% resequenced were chosen. Since DNA samples from only 23 European derived individuals were harvested for SNPs, we checked that the use of $r^2$ to estimate SNP coverage would be robust and not biased due to sampling error. Although this was previously shown to be satisfactory using a coalescence model (5), we used genotype data from six genomic regions from our own laboratory, derived from between 182 and 944 unrelated individuals, consisting of 213 SNPs, and spanning a total of 338 kb (Supplementary Material, Table 2). We confirmed that 23 individuals are sufficient for $r^2$ estimations for our purpose, provided that the $r^2$ values of interest are greater than 0.5 (Supplementary Material Figures).

**Marker sampling simulations**

For each set of simulations a pre-determined target marker density was set (e.g. one SNP per 5 kb). The number of markers sampled was the minimum number of markers required to achieve the target density. If a gene segment was less polymorphic than a target density, then all SNPs in that segment were selected. Three sampling strategies were used: (i) random sampling from the pool of all genes; (ii) random sampling within each gene to the given density; and (iii) sampling with the aim of even spacing by choosing each sampled to be as close to a pre-set evenly spaced position as possible. Ten thousand sampling trials were taken for the random sampling methods for each target density.

**Calculation of LD statistic**

The indirect ascertainment of unobserved SNPs was calculated using the pair-wise LD measurement $r^2$ (17). This LD statistic is inversely related to the required sample size for association mapping, given a fixed genetic effect (18,19). An unobserved SNP was regarded as captured by an observed one, if their pair-wise $r^2$ reached the threshold of interest. We chose to focus on SNP ascertainment instead of LD statistics per se as the former is more directly informative for studying disease association.

**htSNP selection**

For each gene, starting from the set of all mapped SNPs, we minimized the SNPs to retain only one variant from each pair with $r^2 \geq 0.8$. This was performed using a ‘backward selection’ procedure.

All sampling and calculations were performed using programs written in C.

**SUPPLEMENTARY MATERIAL**

Supplementary Material is available at HMG Online.

**ACKNOWLEDGEMENTS**

We would like to thank S. Nezhentsev, B. Barratt and R. Twells for part of the raw data used in Supplementary Material Table 2, and H. Cordell, D. Clayton, J. Cooper and other members of our statistics group for their advice and encouragement. W.Y.S.W. is a recipient of scholarships from the University of Cambridge Clinical School, Gonville and Caius College, and the University of Sydney. This work was funded by the Wellcome Trust and the Juvenile Diabetes Research Foundation International.

**REFERENCES**

11. Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D.,
  Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L. et al.
  (2001) A map of human genome sequence variation containing 1.42 million
  Blocks of limited haplotype diversity revealed by high-resolution scanning
13. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. and Lander, E.S.
  Genet., 29, 229–232.
14. Phillips, M.S., Lawrence, R., Sachidanandam, R., Morris, A.P., Balding, D.J.,
  et al. (2003) Chromosome-wide distribution of haplotype blocks and the
  Detecting disease associations due to linkage disequilibrium: a
  class of tests and the determinants of statistical power. Hum. Hered.
  (in press).
  mating populations. Heredity, 33, 229–239.
  linkage versus association analysis of quantitative traits, by use of
  variance-components models, for sibship data. Am. J. Hum. Genet.,
  66, 1616–1630.