European Journal of Human Genetics (2004) 12, 259–262

www.nature.com/ejhg

LETTER

Is haplotype tagging the panacea to association mapping studies?

Ansar Jawaid, Pak C Sham, Andrew J Makoff, Philip J Asherson*

MRC Social Genetic Developmental Psychiatry Research Centre (SGDP), Institute of Psychiatry, Kings College London, UK

European Journal of Human Genetics (2004) **12,** 259–262. doi:10.1038/sj.ejhg.5201146
Published online 21 January 2004

It is commonly believed that creating population-specific high-density linkage disequilibrium (LD) maps is an important step towards an efficient and cost-effective approach to scanning the genome for associations.[1–5] The basis for this assertion is that single-marker LD mapping methods are unable to describe disequilibrium across chromosomal regions that surround susceptibility loci. It is therefore suggested that a subset of SNPs, which tag common haplotypes describing most genetic diversity, are identified and combined into multimarker haplotypes. This, it is argued, will provide a more cost-efficient approach by reducing the amount of genotyping, in comparison to single-locus studies. This view has led to the launch of a large publicly funded effort to generate such maps in four different population samples (the HAPMAP project), which will become a major international resource (http://www.genome.gov/research).

This raises several issues for those engaged in mapping genes for common disorders, not least the marker density required to define haplotypes blocks accurately, how best to make use of LD map data and the most efficient method for genotyping. Even when sets of haplotype tagging SNPs (htSNPs) are identified, the need to genotype hundreds of individuals for thousands of markers remains prohibitively expensive for most investigators using currently available methods that depend upon PCR. For this reason DNA pooling, in which individual samples of DNA are combined together in pools from which allele frequencies are estimated, is widely recognised as a powerful method that substantially reduces the cost and feasibility of large-scale association studies.[6–16] However, haplotype mapping, in contrast to single-marker analysis, generally requires individual samples to be genotyped. To address this issue, we have examined the informativeness of haplotype mapping with htSNPs in comparison to single-locus analyses using the data published by Johnson *et al*.[2]

In this paper, the authors scanned 135 kb of DNA in nine genes for polymorphic variation, using denaturing high-performance liquid chromatography followed by sequence analysis. They identified and subsequently genotyped 115 polymorphic SNPs with an average spacing of 1174 bp determining haplotypes for each gene in 384 Caucasians of European descent. These data show that for common haplotypes, 34 SNPs retain most of the LD information between adjacent markers.

We have examined the merits of using these 34 htSNPs to:(a) capture the information of the tagged haplotype (*haplotype approach*) and (b) use as markers for single-locus analyses (*tag approach*). We have also selected SNPs on the basis of their minor allele frequencies (MAFs) using three cutoffs: (c) $MAF \geq 20\%$, (d) $MAF \geq 10\%$, and (e) $MAF \geq 5\%$. We tested the relative efficiency of the five strategies to detect associations with each of the 115 markers in turn. This provides a test of the relative power of each approach to detect disease associations assuming that any of the 115 SNPs could be a functionally significant variant (FSV) causing disease susceptibility.

Johnson *et al*[2] provide the frequencies of haplotypes across the nine genes. This enabled us to derive the MAFs for each of the 115 SNPs. For each of the five strategies, we examined the strength of association between the markers selected for use in each strategy and individual markers (representing putative FSVs) across each corresponding gene region. This was carried out using standard contingency χ^2 tests, $2 \times h$ tables in the case of the haplotype analysis approach (h being the number of haplotypes) and a series of 2×2 tables for the single-marker analytic

***Correspondence: Dr Philip J Asherson, Social Genetic Developmental Psychiatry Research, Institute of Psychiatry, Kings College, London, UK. Tel: +44 207 848 0078; Fax: +44 207 848 0407;**
E-mail: p.asherson@iop.kcl.ac.uk
Received 17 March 2003; revised 12 October 2003; accepted 30 October 2003

Table 1 Comparison of the efficiency of the haplotype and single-locus approaches to detect each of the total number of 115 SNPs with $\geq$80% information retained

FSV allele frequency (%)	*Number of markers in allele frequency category*	*Haplotype approach*	*Tag approach*	*MAF $\geq$20% approach*	*MAF $\geq$10% approach*	*MAF $\geq$5% approach*
<5	25	6 (23)	1 (4)	0	0	0
5–19	55	54 (98)	40 (73)	8 (15)	40 (73)	55 (100)
20–49	35	29 (86)	20 (57)	35 (100)	35 (100)	35 (100)
ALL	115	89 (77)	61 (53)	43 (48)	75 (65)	90 (78)
$\overline{\chi}^2$		79	63	49	66	77

The highest χ^2 value for association with the FSV, achieved from the appropriate selection method, is divided by the χ^2 statistic obtained by the FSV itself (perfect association). The efficiency of each strategy is measured as the proportion of the FSVs where the ratio of χ^2 statistics is greater than 80%. Numbers of markers detected in each category (%) with in parentheses. The bottom row of the table shows the mean χ^2 for each approach for all 115 markers ($\overline{\chi}^2$).

strategies. The largest χ^2 statistic for the association of the selected markers with the FSV was recorded and divided by the χ^2 statistic obtained by the FSV with itself, that is, perfect association. We measured the efficiency of each strategy as the proportion of the FSVs where the ratio of χ^2 statistics was greater than 80%. To simplify the presentation of the data, we categorised FSVs on the basis of their MAFs, corresponding to <5, 5–19 and 20–49%.

As shown in Table 1, the use of htSNPs when combined in multipoint haplotype analyses is far better than using the same htSNPs or MAF $\geq$20% SNPs in single-locus tests of associations. While the haplotype approach is still better than the use of MAF $\geq$10% SNPs (77 *versus* 65%), the use of MAF $\geq$5% SNPs is almost identical (78%). When the ability to detect the higher frequency FSVs is compared by combining the 5–19 and 20–50% FSV groups, the MAF $\geq$5% approach is the most efficient (100%) followed by the haplotype approach (92%) and the MAF $\geq$10% approach (83%). The use of DNA pooling to screen MAF $\geq$5% SNPs therefore compares favourably with the haplotype approach for detecting common FSVs.

The DNA pooling approach outlined above does, however, require rescreening of coding and noncoding functional genomic regions to provide sufficiently comprehensive maps of polymorphic markers. The reason for this is that the current generation of SNP maps are far from adequate for the DNA pooling approach, despite the fact that at the time of writing 3 049 569 reference SNPs are listed in dbSNP at the National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov/SNP/), 522 072 of which have been validated. Of these, it is estimated that approximately 50% of the candidate SNPs have common minor alleles with frequency greater than 20% in any given population, approximately 20% have less common alleles with frequency between 5 and 20% and the rest are accounted for by rare variants ($<$5%) and sequencing artefacts.[17] Johnson *et al*[2] note that only 25% of the htSNPs they identified were found on the database, which is in fact not surprising since in their data only 10 out of 34 htSNPs had MAFs greater than 20%. Furthermore, although currently available reference SNPs have an average spacing of around one every 1 kb, there are many gaps in the map and many of the SNPs do not lie within or close to likely functional regions. For example, we have examined the available SNPs for 12 ion-channel genes that span a total of 1317 kb of DNA. Out of a total number of 661 listed SNPs, we found 120 gaps of 2–5 kb, 43 gaps of 5–10 kb and 20 gaps of $>$10 kb, despite an average map density of 2.2 kb for these genes.

What is the feasibility of performing the large number of individual SNP assays required for the haplotype approach? If we consider a gene of average size 100 kb, we would expect around 200 SNPs with allele frequency $>$5%[18] and 50 htSNPs assuming the proportion of htSNPs observed by Johnson and co-workers. If we were to screen 1000 genes in 500 cases and 500 controls, we would need to perform 50 000 000 individual genotypes to adopt the haplotype approach. In contrast, if we adopted a DNA pooling approach using quadruplicate assays in two pools as outlined in the recent review by Sham *et al*,[16] we would require 1 600 000 SNP genotypes.

The HAPMAP project that aims to develop genome-wide haplotype maps is taking a different strategy to the one proposed for DNA pooling, since it is not planned to base these on comprehensive SNP maps. Although a comprehensive genome-wide map and haplotype analysis of all genetic variation is the ideal, this would require a considerable amount of additional work to generate such data and would be hugely expensive. It is in fact estimated that generating the haplotype map will require successful genotyping of 450 000 SNPs (around one SNP every 6–7 kb) and will require initial testing of some 800 000 to 900 000 SNPs (http://www.genome.gov/research). Haplotype maps based on such data have been shown to generate SNPs that

tag common haplotypes that account for a large proportion of genetic variation, so it should be possible to use HAPMAP markers to screen the genome for associations to common haplotypes.[1] However, it is not yet known how the efficiency of a strategy based on incomplete SNP maps will compare to strategies based on complete information such as the data from Johnson *et al*.[2] Evidence that the current maps are not sufficiently dense for this purpose comes from the recent analysis of a first-generation haplotype map of chromosome 19.[19] Using publicly available SNPs, the authors show that one-third of the chromosome is encompassed within haplotype blocks. However, evolutionary modelling of the data indicate the dependence of observed block lengths on marker spacing and allele frequency, suggesting that apparent blocks can stem from incomplete coverage of the chromosome genealogy. They conclude that genotyping additional markers is likely to uncover further recombination events, breaking up larger blocks and refining their boundaries.

Another important unanswered question in the use of haplotype maps is whether the data generated in one sample population will generalise to other sample populations. If we depend on a single central haplotype map for broadly defined population groups such as Caucasians, Asians or Africans, the robustness of the approach will crucially depend on how comparable the haplotype map is to the populations from which samples for association studies are being drawn. While comparable data have been described across populations, the true extent of differences in haplotype frequencies in apparently related populations and the influence that such differences will have on the power of the analyses remains unclear.

Finally, it may be possible to combine DNA pooling and htSNP approaches by using DNA pooling data to derive common haplotype frequencies. Clayton *et al*[15] have demonstrated that frequencies of common haplotypes can be estimated using DNA pooling data on selected htSNPs in the light of good prior knowledge of haplotype structure. Although such an approach was possible using the data from Johnson *et al*[2] based on a complete SNP map of the regions investigated, it remains unknown how often such situations will occur in practice and whether the approach can be generalised to a HAPMAP based on incomplete SNP data.

Despite the utility of DNA pooling approaches, some limitations remain (reviewed in Sham *et al*[16]). The method is more easily applied to the analysis of categorical data using a simple case–control design, although it is possible perform within family tests of association by comparing proband pools to parent pools. Other strategies enable the analysis of quantitative phenotypes by comparing pools representing high and low scoring groups (eg comparison of top and bottom decile groups for normally distributed traits) or comparison of intermediate groups to detect genotype–phenotype correlations across the distribution (eg comparison of pools for all 10 decile groups). Within sibling-pair tests of association that are robust to stratification can be performed by comparing pools of low scoring siblings with pools of their high scoring cosiblings (ie a discordant sibling test of association). A limited number of covariates can be included by comparing pools grouped by covariate in addition to phenotypic score in a 2×2 pooling design (eg high and low depression scores, with and without adverse life events). DNA pooling and htSNP approaches are both powerful methods for testing the *common disease common variant* hypothesis, but may have limited utility in the face of allelic heterogeneity. Finally, it should be recognised that while DNA pooling is an effective tool for cost-effective and rapid screening for SNP associations, final analysis requires individual genotyping to clarify the potential role of specific functional variants or functionally significant haplotypes where SNPs may act in combination to produce disease susceptibility.

In conclusion, we suggest that at this time common genetic variants that confer risk to common complex disorders (the 'low-hanging fruit') will be more efficiently detected through the adoption of DNA pooling strategies based on comprehensive SNP maps of targeted functional regions. In agreement with Johnson *et al*,[2] we conclude that current SNP databases may have limited utility for LD mapping because it may not be possible to define many common haplotypes using incomplete maps and that a directed resequencing effort of approximately 10% of the genome in or near genes in the major ethnic groups is required for a complete evaluation of the common variant model. Haplotype maps are likely to have a greater impact when the issues around map density, general applicability between sample populations and cost-effective processing of large numbers of genotypes have been resolved.

References

1 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the genome. *Science* 2002; **296**: 2225–2229.
2 Johnson GC, Esposito L, Barratt BJ *et al*: Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001; **29**: 233–237.
3 Goldstein DB: Islands of linkage disequilibrium. *Nat Genet* 2001; **29**: 109–111.
4 Reich DE, Cargill M, Bolk S *et al*: Linkage disequilibrium in the human genome. *Nature* 2001; **411**: 199–204.
5 Stephens JC, Schneider JA, Tanguay DA *et al*: Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 2001; **293**: 489–493.
6 Sasaki T, Tahira T, Suzuki A *et al*: Quantitative SSCP analysis of pooled DNA. *Am J Hum Genet* 2001; **68**: 214–218.
7 Wolford JK, Blunt D, Ballecer C, Prochazka M: High-throughput SNP detection by using denaturing high performance liquid chromatography (DHPLC). *Hum Genet* 2000; **107**: 483–487.
8 Breen G, Harold D, Ralston S *et al*: Determining SNP allele frequencies in DNA pools. *Biotechniques* 2000; **28**: 464–470.
9 Hoogendoorn B, Norton N, Kirov G *et al*: Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Hum Genet* 2000; **107**: 488–493.

10 Curran S, Hill L, O'Grady G *et al*: Validation of single nucleotide polymorphism quantification in pooled DNA samples with SNaPIT – a glycosylase-mediated methods for polymorphism detection method. *Mol Biotechnol* 2002; **22**: 253–262.
11 Jawaid A, Bader JS, Purcell S *et al*: Optimal selection strategies for QTL mapping using pooled DNA samples. *Eur J Hum Genet* 2002; **10**: 125–132.
12 Barcellos LF, Klitz W, Field LL *et al*: Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am J Hum Genet* 1997; **61**: 734–747.
13 Shaw SH, Carrasquillo MM, Kashuk C *et al*: Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Res* 1998; **8**: 111–123.
14 Risch N, Teng J: The relative power of family-based and case–control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res* 1999; **8**: 1273–1288.
15 Barratt BJ, Payne F, Rance HE *et al*: Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann Hum Genet* 2002; **66**: 393–405.
16 Sham P, Bader JS, Craig I *et al*: DNA Pooling: a tool for large-scale association studies. *Nat Rev Genet* 2002; **3**: 862–871.
17 Marth G, Yeh R, Minton M *et al*: Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat Genet* 2001; **27**: 371–372.
18 Phillips MS, Lawrence R, Sachidanandam R *et al*: Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat Genet* 2003; **33**: 382–387.
19 Kruglyak L, Nickerson DA: Variation is the spice of life. *Nat Genet* 2001; **27**: 234–236.