

ARTICLE

# Validity of tagging SNPs across populations for association studies

Albert Tenesa<sup>\*</sup>,<sup>1,2</sup> and Malcolm G Dunlop<sup>1,2</sup>

<sup>1</sup>Colon Cancer Genetics Group, University of Edinburgh, School of Molecular and Clinical Medicine, Western General Hospital, Crewe Road, Edinburgh, UK; <sup>2</sup>MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh, UK

Recent advances in high throughput genotyping technologies will allow large-scale association studies to disentangle the genetic basis of human common diseases. Currently, a large-scale genotyping effort is being carried out by the HapMap project and the outcome of this project is expected to help researchers in their efforts to understand how genetic variation influences susceptibility to disease. However, there is some controversy on whether this huge public effort will be of value for those populations not studied in the HapMap project. Here, we present simulation results based on the empirical distribution of linkage disequilibrium (LD) on a large chromosomal region (10 Mb) on human chromosome 20<sup>1,2</sup> for two European and two Asian populations. These results show that statistical power to detect associations does not depend on the population where SNP tagging was performed.

*European Journal of Human Genetics* (2006) 14, 357–363. doi:10.1038/sj.ejhg.5201554; published online 4 January 2006

**Keywords:** LD; HapMap; SNP; tSNP; mapping

## Introduction

In an ideal situation, researchers involved in mapping variants predisposing to disease would study most of the genetic variation present in the population by resequencing a representative sample of cases and controls from the same population.<sup>1,2</sup> Since polymorphisms that predispose to human disease are expected to be of small effect<sup>3</sup> large-scale association studies are required to detect such effects. Hence, resequencing is at the moment impracticable on such large scale, and so researchers working on large-scale association studies rely on the correlation among linked loci (linkage disequilibrium (LD)) to find those polymorphisms that predispose to human disease. Since LD is a population-dependent parameter, there has been controversy on whether LD patterns observed in one population

would be observed in a different population and therefore doubts have been expressed on the utility of large-scale genotyping efforts such as the HapMap project.<sup>4,5</sup> The utility of such an effort for mapping studies will ultimately depend on the power the subset of loci typed on the study population has to detect the disease locus (DL) or loci and not necessarily on whether LD patterns or block boundaries are conserved among populations. Here, we present a simulation study in which SNPs tagging (tSNPs) a 10 Mb region of human chromosome 20 were selected independently for each population and tested for association both within the same population and also on an independent population. Power (defined as the proportion of all tSNPs that showed a significant association with disease, at the 5% level) did not depend on which population was used for tagging.

\*Correspondence: Dr A Tenesa, Colon Cancer Genetics Group, School of Molecular and Clinical Medicine, The University of Edinburgh, 4th Floor, MRC Human Genetics Unit, Western General Hospital, Crewe Road South, Edinburgh EH4 2XU, UK. Tel: +44 131 332 2471; Fax: +44 131 467 8450; E-mail: albert.tenesa@ed.ac.uk  
Received 22 April 2005; revised 3 November 2005; accepted 10 November 2005; published online 4 January 2006

## Materials and methods

### Data

We used available genotype data for a 10 Mb region on chromosome 20 (<http://www.well.ox.ac.uk/~xiayi/data/chr20/10Mb/index.html>). A total of 4427, 5324, 4160 and 4160 SNPs were available for  $n=96$ , 47, 20 and 22

unrelated individuals from four different populations (UK Caucasian, CEPH, Han Chinese and Japanese, respectively). For each population pair, we first selected those loci that were segregating in the two populations and did not have any missing values. The number of loci analysed for each population comparison varied from 1012 (for UK Caucasian and CEPH) to 2100 (for Han Chinese and Japanese).

### Tagging methods

**SpD** tSNPs were selected using the *spD* method described by Meng *et al.*<sup>6</sup> This method is based on calculating the pair-wise LD matrix among  $m$  candidate markers where the components of the  $m \times m$  matrix are the composite LD measure<sup>7</sup> divided by its SD. Standard principal components analysis is performed on this matrix and those SNPs that contribute most to the eigenvectors of the largest eigenvalues are selected as tags.

For large genomic regions as the one considered here, the method is applied on sliding windows of much smaller size and SNPs selected as tSNPs if they had been selected in a given proportion of the windows in which they were included.

tSNPs were selected using 20 individuals with a sliding window size of five SNPs. The selected SNPs explained 85% of the variation within each sliding window and were finally selected if they were preselected in 80% of the

sliding windows in which they were present. These same parameters were used for all the results concerning the *SpD* tagging method. The results shown below are based on *SpD* unless otherwise stated.

**ldSelect** In order to investigate the effect of the tagging method on the portability of tSNPs across populations *ldSelect*<sup>8</sup> was also used to select tSNPs. As with *SpD*, 20 individuals were selected with replacement among the  $n$  available individuals to be used as the selection panel. The program was then run with the following parameters:  $r^2$  value<sup>9</sup> of 0.8 and no selection of SNPs by allele frequency.

### Simulations

After selecting tSNPs, case-control samples were generated by first selecting one of the available SNPs at random and assuming it to be the DL. The DL could be a tSNP or not. Then, one of the two alleles at the DL was selected at random to be the one that increased disease susceptibility. Finally, a sample of  $N$  cases and controls was generated as follows:

**Simulation of cases** Assume  $Q$  is the allele that increases disease susceptibility at the DL. Genotype frequencies at the DL were estimated from the  $n$  available individuals.

**Table 1** Across replicates mean proportion of SNPs selected as tags (*SpD* tagging method), the mean and the SD of the proportion (over the total of tSNPs) of the tSNPs that were significant at the 5% level

Comparison	tSNP population	Case-control population	Mean proportion of markers selected as tags	Mean proportion of significant tSNPs	SD proportion of significant tSNPs
1	UK Caucasian	UK Caucasian	0.42	0.23	0.09
	Japanese	UK Caucasian	0.41	0.23	0.10
	UK Caucasian	Japanese	0.44	0.49	0.14
	Japanese	Japanese	0.39	0.48	0.15
2	UK Caucasian	UK Caucasian	0.48	0.22	0.10
	CEPH	UK Caucasian	0.47	0.22	0.09
	UK Caucasian	CEPH	0.48	0.34	0.13
	CEPH	CEPH	0.47	0.34	0.13
3	CEPH	CEPH	0.39	0.34	0.12
	Japanese	CEPH	0.38	0.35	0.13
	CEPH	Japanese	0.39	0.49	0.15
	Japanese	Japanese	0.36	0.48	0.14
4	UK Caucasian	UK Caucasian	0.44	0.22	0.10
	Han Chinese	UK Caucasian	0.42	0.22	0.10
	UK Caucasian	Han Chinese	0.44	0.50	0.15
	Han Chinese	Han Chinese	0.42	0.49	0.14
5	CEPH	CEPH	0.39	0.35	0.13
	Han Chinese	CEPH	0.39	0.35	0.13
	CEPH	Han Chinese	0.39	0.51	0.14
	Han Chinese	Han Chinese	0.38	0.50	0.15
6	Han Chinese	Han Chinese	0.34	0.50	0.14
	Japanese	Han Chinese	0.33	0.51	0.14
	Han Chinese	Japanese	0.34	0.50	0.13
	Japanese	Japanese	0.34	0.49	0.14

**Table 2** Across replicates mean proportion of SNPs selected as tags (*SpD* tagging method), the mean and the SD of the proportion of tSNPs (over the total of tSNPs) that were significant at the 5% level

Comparison	tSNP population	Case-control population	Mean proportion of markers selected as tags	Mean proportion of significant tSNPs	SD proportion of significant tSNPs
1	UK Caucasian	UK Caucasian	0.38	0.47	0.16
	Japanese	UK Caucasian	0.41	0.47	0.17
	UK Caucasian	Japanese	0.41	0.49	0.15
	Japanese	Japanese	0.40	0.49	0.15
2	UK Caucasian	UK Caucasian	0.42	0.49	0.16
	CEPH	UK Caucasian	0.47	0.48	0.17
	UK Caucasian	CEPH	0.43	0.51	0.16
	CEPH	CEPH	0.46	0.50	0.16
3	CEPH	CEPH	0.37	0.48	0.15
	Japanese	CEPH	0.37	0.49	0.15
	CEPH	Japanese	0.37	0.49	0.14
	Japanese	Japanese	0.37	0.49	0.15
4	UK Caucasian	UK Caucasian	0.38	0.47	0.17
	Han Chinese	UK Caucasian	0.42	0.47	0.17
	UK Caucasian	Han Chinese	0.38	0.50	0.14
	Han Chinese	Han Chinese	0.42	0.50	0.14
5	CEPH	CEPH	0.34	0.50	0.15
	Han Chinese	CEPH	0.38	0.50	0.16
	CEPH	Han Chinese	0.35	0.50	0.15
	Han Chinese	Han Chinese	0.36	0.50	0.14
6	Han Chinese	Han Chinese	0.33	0.50	0.13
	Japanese	Han Chinese	0.33	0.50	0.14
	Han Chinese	Japanese	0.34	0.50	0.13
	Japanese	Japanese	0.33	0.49	0.14

5000 cases and controls were typed. Incidence was 0.005 and genotype relative risks were 1,  $\sqrt{2.5}$  and 2.5. Here the number of initial people genotyped is kept constant for all populations ( $n = 20$ ).

The prevalence of the disease in the population ( $K$ ) is equal to

$$K = P(D|QQ) \times P(QQ) + P(D|Qq) \times P(Qq) + P(D|qq) \times P(qq)$$

where  $P(D|G)$  is the penetrance and  $P(G)$  is the frequency of the  $G$  genotype.

The genotype relative risks for  $QQ$ ,  $Qq$  and  $qq$  genotypes are respectively  $GRR_{QQ} = P(D|QQ)/P(D|qq)$ ,  $GRR_{Qq} = P(D|Qq)/P(D|qq)$  and  $GRR_{qq} = 1$ .

Then,

$$P(D|qq) = K / [GRR_{QQ} \times P(QQ) + GRR_{Qq} \times P(Qq) + GRR_{qq} \times P(qq)]$$

$$P(D|Qq) = P(D|qq) \times GRR_{Qq}$$

$$P(D|QQ) = P(D|qq) \times GRR_{QQ}$$

Case subjects were simulated by selecting individuals at random, checking the genotype ( $G$ ) at the preassigned DL and drawing a random number between 0 and 1 from a uniform distribution. If this number was equal or less than  $P(D|G)$  then the individual was considered a case if it was

not the individual was rejected and a new individual drawn at random. The process was repeated until  $N$  cases were obtained.

**Simulation of controls** It was assumed that the genotype frequency of the control subjects was the same as the general population. Therefore, controls were simulated by selecting with replacement  $N$  individuals from the  $n$  available samples.

**Replicates** In order to account for:

- (1) Variation among samples used for tagging, 10 different samples were obtained by sampling with replacement 20 individuals among the  $n$  individuals from each population.
- (2) Variation due to different case-control samples was taken into account by sampling with replacement four replicates within each tSNP sample.
- (3) Differences in LD patterns among different regions within the 10 Mb region, 100 different loci were simulated within each case-control sample.

**Table 3** The mean and the SD of the proportion of tSNPs (over the total of tSNPs) that were significant at the 5% level when using *ldSelect*

Comparison	tSNP population	Case-control population	Mean (SD) overall	Mean (SD) DL is tSNP	Mean (SD) DL not tSNP
1	UK Caucasian	UK Caucasian	0.15 (0.06)	0.15 (0.06)	0.16 (0.05)
	Japanese	UK Caucasian	0.15 (0.06)	0.14 (0.06)	0.15 (0.06)
	UK Caucasian	Japanese	0.35 (0.11)	0.34 (0.11)	0.38 (0.10)
	Japanese	Japanese	0.35 (0.11)	0.35 (0.11)	0.37 (0.10)
2	UK Caucasian	UK Caucasian	0.14 (0.06)	0.14 (0.06)	0.16 (0.05)
	CEPH	UK Caucasian	0.15 (0.06)	0.14 (0.06)	0.16 (0.05)
	UK Caucasian	CEPH	0.22 (0.08)	0.22 (0.08)	0.25 (0.08)
	CEPH	CEPH	0.22 (0.08)	0.22 (0.08)	0.25 (0.08)
3	CEPH	CEPH	0.23 (0.08)	0.23 (0.08)	0.23 (0.08)
	Japanese	CEPH	0.23 (0.08)	0.23 (0.08)	0.25 (0.08)
	CEPH	Japanese	0.34 (0.11)	0.34 (0.11)	0.37 (0.10)
	Japanese	Japanese	0.34 (0.11)	0.34 (0.11)	0.37 (0.09)
4	UK Caucasian	UK Caucasian	0.15 (0.06)	0.15 (0.06)	0.16 (0.06)
	Han Chinese	UK Caucasian	0.15 (0.06)	0.15 (0.06)	0.15 (0.06)
	UK Caucasian	Han Chinese	0.36 (0.11)	0.35 (0.11)	0.37 (0.11)
	Han Chinese	Han Chinese	0.36 (0.11)	0.35 (0.11)	0.38 (0.10)
5	CEPH	CEPH	0.23 (0.08)	0.22 (0.08)	0.25 (0.08)
	Han Chinese	CEPH	0.23 (0.08)	0.23 (0.09)	0.24 (0.08)
	CEPH	Han Chinese	0.35 (0.11)	0.35 (0.12)	0.37 (0.11)
	Han Chinese	Han Chinese	0.36 (0.11)	0.36 (0.11)	0.38 (0.09)
6	Han Chinese	Han Chinese	0.37 (0.11)	0.36 (0.11)	0.38 (0.10)
	Japanese	Han Chinese	0.36 (0.11)	0.36 (0.11)	0.38 (0.11)
	Han Chinese	Japanese	0.34 (0.11)	0.34 (0.11)	0.36 (0.10)
	Japanese	Japanese	0.35 (0.11)	0.34 (0.12)	0.37 (0.10)

5000 cases and controls were typed. Incidence was 0.005 and genotype relative risks were 1,  $\sqrt{2.5}$  and 2.5. Results are shown for three different scenarios: overall (it is not known whether the DL is a tSNP or not), DL is a tSNP and DL is not a tSNP.

In summary, we used resampling methods within one population in order to investigate the behaviour of tags within that population. We compared these results to those obtained by testing those tags in different populations. If similar results were obtained from tags developed in one population (tagging population) and applied to the second population (ie case-control population), then these simulations would suggest that tSNPs would have validity between populations.

## Results

Table 1 shows the average power obtained for a case-control study with 5000 cases and 5000 controls when the assumed disease model is multiplicative (ie genotype relative risks were 1,  $\sqrt{2.5}$  and 2.5) and the prevalence was equal to 0.005.

The percentage of tSNPs that showed statistical significance in a case-control study did not substantially vary with the population on which the tagging was performed.

The higher percentage of significant tSNPs observed for the Han Chinese and Japanese compared to UK Caucasian and to a lesser extent to CEPH reflects the fact that there is

a much smaller sample size ( $n$ ) for those two populations and cases and controls tend to be more alike increasing therefore the chances of detecting an association. In order to investigate this further,  $n = 20$  individuals for each of the populations were selected at random without replacement and the same analysis performed. Table 2 shows that if  $n$  is kept equal in all populations then this effect disappears.

We used a different tagging method (*ldSelect*) to assess the effect the tagging method has on the portability of tSNPs across populations. *ldSelect*-selected tags also showed that tSNPs selected in one population are valid for another (Table 3). The different proportion of significant tSNPs for *SpD* and *ldSelect* is because the proportion of variation explained by the tSNPs is not exactly comparable for the two methods. Table 3 shows that there are small differences when the DL is selected as tSNP or not.

A higher proportion of tSNPs show a significant association when the DL is not a tSNP than when it is. This is probably because more than one tSNPs are selected to tag the DL when the DL is not a tag, whereas only one (the DL) is required when the DL is a tSNP.

Table 4 shows results for three different scenarios (A) DL is a tSNP (B) DL is not a tSNP but is in the original set used for tSNP selection (C) DL is not in the original set for tSNP

**Table 4** The mean and the SD of the proportion of tSNPs (over the total of tSNPs) that were significant at the 5% level when using *SpD*

Comparison	tSNP population	Case-control population	Scenario A mean (SD)	Scenario B mean (SD)	Scenario C mean (SD)
1	UK Caucasian	UK Caucasian	0.08 (0.04)	0.08 (0.04)	0.08 (0.04)
	Japanese	UK Caucasian	0.08 (0.04)	0.09 (0.04)	0.08 (0.04)
	UK Caucasian	Japanese	0.24 (0.11)	0.24 (0.11)	0.24 (0.11)
	Japanese	Japanese	0.23 (0.11)	0.25 (0.11)	0.25 (0.11)
2	UK Caucasian	UK Caucasian	0.08 (0.04)	0.09 (0.04)	0.09 (0.04)
	CEPH	UK Caucasian	0.08 (0.04)	0.09 (0.04)	0.08 (0.04)
	UK Caucasian	CEPH	0.14 (0.07)	0.15 (0.07)	0.14 (0.07)
	CEPH	CEPH	0.13 (0.07)	0.14 (0.07)	0.14 (0.08)
3	CEPH	CEPH	0.14 (0.07)	0.15 (0.07)	0.14 (0.07)
	Japanese	CEPH	0.14 (0.07)	0.15 (0.07)	0.15 (0.08)
	CEPH	Japanese	0.23 (0.11)	0.25 (0.11)	0.22 (0.11)
	Japanese	Japanese	0.23 (0.11)	0.25 (0.11)	0.23 (0.10)
4	UK Caucasian	UK Caucasian	0.08 (0.04)	0.09 (0.04)	0.09 (0.04)
	Han Chinese	UK Caucasian	0.08 (0.04)	0.09 (0.04)	0.09 (0.04)
	UK Caucasian	Han Chinese	0.24 (0.12)	0.25 (0.12)	0.24 (0.12)
	Han Chinese	Han Chinese	0.24 (0.11)	0.27 (0.12)	0.27 (0.11)
5	CEPH	CEPH	0.13 (0.07)	0.15 (0.07)	0.15 (0.07)
	Han Chinese	CEPH	0.15 (0.08)	0.15 (0.07)	0.14 (0.07)
	CEPH	Han Chinese	0.24 (0.12)	0.26 (0.12)	0.25 (0.12)
	Han Chinese	Han Chinese	0.23 (0.12)	0.26 (0.12)	0.23 (0.11)
6	Han Chinese	Han Chinese	0.24 (0.12)	0.26 (0.12)	0.25 (0.13)
	Japanese	Han Chinese	0.25 (0.13)	0.26 (0.12)	0.27 (0.12)
	Han Chinese	Japanese	0.23 (0.11)	0.26 (0.11)	0.24 (0.11)
	Japanese	Japanese	0.23 (0.11)	0.24 (0.11)	0.24 (0.11)

Incidence was 0.01 and genotype relative risks were 1, 1.7 and 2.9. Results are shown for three different scenarios (A) DL is a tSNP (B) DL is not a tSNP but is in the original set used for tSNP selection and (C) DL is not in the original set for tSNP selection. In all, 1000 cases and controls were typed.

selection. Differences in power are very small under the three scenarios. As shown above for *ldSelect* scenario B has a slight increase in the proportion of significant tSNPs compared to scenario A. Similarly, for scenarios C and A.

We repeated the analysis but this time we split the SNPs into two categories: (A) those SNPs that were within genes and (B) those SNPs that were outwith genes. Table 5 shows a summary of these results. Overall, SNPs within genes exhibited a higher proportion of significant tSNPs, although the difference was very small. The population from which tSNPs were selected had little or no effect on the proportion of significant tSNPs both for SNPs within and out of genes.

Figure 1 shows the fitted second-degree polynomial to the percentage of significant tSNPs as a function of the DL frequency when the tagging and the case-control study was performed on the same or different population. Lines mostly overlap over the whole DL frequency spectrum studied. Taken together, these results suggest that SNPs tagged by the HapMap project on the CEPH, Han Chinese and Japanese samples would yield similar power in populations of European and Asian descent.

## Discussion

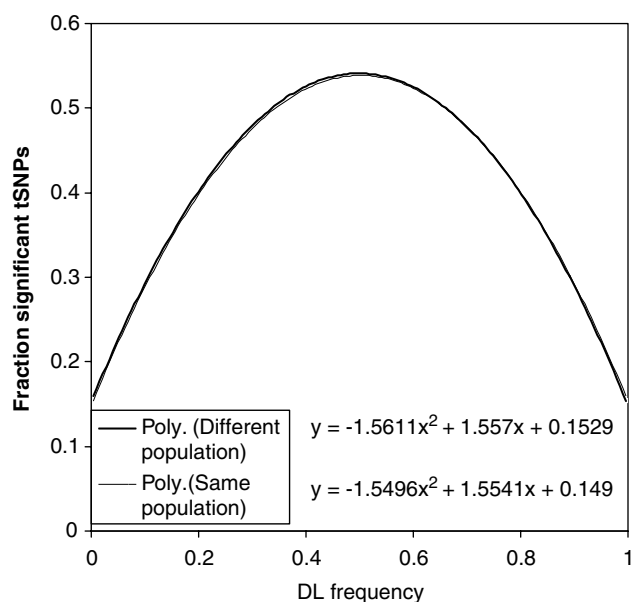
We have shown that tSNPs selected in Han Chinese, Japanese, CEPH and UK Caucasian populations perform similarly (in terms of power) in all these populations.

Ahmadi *et al*<sup>10</sup> studied the performance of tSNPs selected in CEPH samples on a Japanese population by evaluating how well tags represent the variation present in another sample. Their study differs from ours in that they focused on 55 genes involved in drug metabolism and transport whereas we focused on a large chromosomal region. Mueller *et al*<sup>11</sup> studied LD patterns across different European populations and concluded that CEPH-derived tags were of restricted applicability to European populations. However, the comparisons were based on only four genes spanning about 749 kb. Evans and Cardon<sup>12</sup> studied the same samples presented here but used only adjacent markers for their analysis. In addition, all of the previous studies based their evaluation of performance in the LD measure<sup>9</sup>  $r^2$  and Evans and Cardon<sup>12</sup> suggested that the differences they observed might be because of the dependence of  $r^2$  on, for example allele frequencies.<sup>12</sup>

**Table 5** The mean and the SD of the proportion of tSNPs (over the total of tSNPs) that were significant at the 5% level when using *SpD*

Comparison	tSNP population	Case-control population	SNPs within genes mean (SD)	SNPs not in genes mean (SD)
1	UK Caucasian	UK Caucasian	0.24 (0.11)	0.22 (0.10)
	Japanese	UK Caucasian	0.25 (0.10)	0.23 (0.10)
	UK Caucasian	Japanese	0.52 (0.14)	0.48 (0.15)
	Japanese	Japanese	0.50 (0.14)	0.49 (0.14)
2	UK Caucasian	UK Caucasian	0.24 (0.10)	0.21 (0.10)
	CEPH	UK Caucasian	0.24 (0.10)	0.21 (0.10)
	UK Caucasian	CEPH	0.37 (0.13)	0.33 (0.13)
	CEPH	CEPH	0.37 (0.13)	0.32 (0.13)
3	CEPH	CEPH	0.36 (0.12)	0.34 (0.13)
	Japanese	CEPH	0.36 (0.12)	0.35 (0.14)
	CEPH	Japanese	0.51 (0.14)	0.49 (0.15)
	Japanese	Japanese	0.51 (0.13)	0.48 (0.15)
4	UK Caucasian	UK Caucasian	0.24 (0.10)	0.22 (0.10)
	Han Chinese	UK Caucasian	0.24 (0.10)	0.23 (0.10)
	UK Caucasian	Han Chinese	0.53 (0.14)	0.50 (0.15)
	Han Chinese	Han Chinese	0.53 (0.13)	0.49 (0.15)
5	CEPH	CEPH	0.35 (0.12)	0.35 (0.13)
	Han Chinese	CEPH	0.36 (0.12)	0.36 (0.12)
	CEPH	Han Chinese	0.53 (0.15)	0.49 (0.15)
	Han Chinese	Han Chinese	0.52 (0.15)	0.50 (0.15)
6	Han Chinese	Han Chinese	0.50 (0.15)	0.51 (0.14)
	Japanese	Han Chinese	0.50 (0.15)	0.51 (0.14)
	Han Chinese	Japanese	0.50 (0.14)	0.49 (0.15)
	Japanese	Japanese	0.49 (0.14)	0.49 (0.14)

In all, 5000 cases and controls were typed. Incidence was 0.005 and genotype relative risks were 1,  $\sqrt{2.5}$  and 2.5. Results are shown for SNPs within genes and SNPs not in genes (Ensembl v32).



**Figure 1** Proportion of significant tSNPs over all populations as a function of the disease locus frequency when the tagging and the case-control study was done on the same or different population.

In our simulation study only SNPs that were segregating in the two populations were used. This will surely overestimate the performance of across population tags since it could be the case that one DL is segregating in one population but not in another or that one SNP might be a tag in one population and not segregate in another. Nonetheless, we have shown (Figure 1) that our conclusions are consistent even for relatively extreme DL allele frequencies. In any case, if a DL was not segregating in a population, it could not be mapped in this population even if the tags had been obtained from it. If a tSNP selected in one population was not segregating in a case-control population, then there would be a reduction in power since a smaller proportion of the genetic variation would be represented. Researchers using the HapMap data might want to check that the SNPs they have selected as tags are segregating in different populations.

The allele frequency distribution of the unselected SNPs and tSNPs was uniform (results not shown), however if there was not ascertainment bias one would expect an L-shaped distribution. Hence, ascertainment bias may have an important effect on our general conclusions since a uniform allele frequency distribution will tend to

overestimate the amount of LD and therefore decrease the number of tSNPs required. Also, low-frequency variants could be missed if they were segregating at low frequency in one population but not segregating at all in another. Hence, the conclusions drawn here will be more appropriate for common genetic variation as studied by the HapMap project, which will help genetic association studies on different populations to those studied by it.

#### Acknowledgements

We thank Peter Visscher and Naomi Wray for helpful comments on an early version of the manuscript and Ian White and Andrew Carothers for helpful discussions on PCA. The relevant work on genetic susceptibility to colorectal cancer ongoing in these laboratories is funded by a Cancer Research UK Programme Grant (C348/A3758) and by grants from the Scottish Executive Chief Scientist Office (K/OPR/2/2/D333 and CZB/4/94) and by the Medical Research Council (G0000657-53203). AT is funded by Grant C348/A3758.

#### References

- 1 Ke X, Durrant C, Morris AP *et al*: Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Hum Mol Genet* 2004; **13**: 2557–2565.
- 2 Ke X, Hunt S, Tapper W *et al*: The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum Mol Genet* 2004; **13**: 577–588.
- 3 Hirschhorn JN, Daly MJ: Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005; **6**: 95–108.
- 4 Nielsen R, Hubisz MJ, Clark AG: Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* 2004; **168**: 2373–2382.
- 5 Terwilliger JD, Haghghi F, Hiekkalinna TS, Goring HHH: A biased assessment of the use of SNPs in human complex traits. *Curr Opin Genet Dev* 2002; **12**: 726–734.
- 6 Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG: Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am J Hum Genet* 2003; **73**: 115–130.
- 7 Weir BS: *Genetic Data Analysis II*. Sunderland, MA: Sinauer Associates, 1996.
- 8 Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: Selecting a maximally informative set of single-nucleotide polymorphisms for association analysis using linkage disequilibrium. *Am J Hum Genet* 2004; **74**: 106–120.
- 9 Hill WG, Robertson A: Linkage disequilibrium in finite populations. *Theo Appl Genet* 1968; **38**: 226–231.
- 10 Ahmadi KR, Weale ME, Xue ZY *et al*: A single-nucleotide polymorphism tagging set for human drug metabolism and transport. *Nat Genet* 2005; **37**: 84–89.
- 11 Mueller JC, Lohmussaar E, Magi R *et al*: Linkage disequilibrium patterns and tagSNP transferability among European populations. *Am J Hum Genet* 2005; **76**: 387–398.
- 12 Evans DM, Cardon LR: A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *Am J Hum Genet* 2005; **76**: 681–687.