ANNALS OF human genetics

doi: 10.1111/j.1469-1809.2008.00434.x

Investigation into the Ability of SNP Chipsets and Microsatellites to Detect Association with a Disease Locus

D. Curtis[1,*], A. E. Vine[1] and J. Knight[2]

[1] *Centre for Psychiatry, Queen Mary's School of Medicine and Dentistry, London E1 1BB, UK*

[2] *Social Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, De Crespigny Park, London SE5 8AF, UK*

Summary

We wished to investigate the ability of different SNP chipsets to detect association with a disease and to investigate the linkage disequilibrium (LD) relationships between microsatellites and nearby SNPs in order to assess their potential usefulness to detect association.

SNP genotypes were obtained from HapMap and microsatellite genotypes from CEPH. 5000 SNPs were simulated as disease genes which increased penetrance from 0.01 to 0.02 in a sample of 400 cases and 400 controls. The power of flanking SNPs to detect association was tested using sets of 1, 2, 3 or 4 markers analysed with haplotype analysis or logistic regression and using either all HapMap markers or those from the Affymetrix 500K, Illumina 300K or Illumina 550K chipsets. Additionally, LD relationships between 10 microsatellites and SNPs within 2Mb of each other were studied.

The power for one of the markers to detect association at $p = 0.001$ was around 0.4. Power was slightly better for logistic regression than haplotype analysis and for two-marker as opposed to single marker analysis but analysing with larger numbers markers had little benefit. The Illumina 550K marker set was better able to detect association than the other two and was almost as powerful as using all HapMap markers. Microsatellites had detectable LD with only a small number of nearby SNPs and the pattern of LD was very variable.

Available chipsets have quite good ability to detect association although obviously results will be critically dependent on the nature of the genetic effect on risk, sample size and the actual LD relationships of the susceptibility polymorphisms involved. Microsatellites seem ill-suited for systematic studies to detect association.

Keywords: Genetic association, simulation, SNP, microsatellite

Introduction

Many previous investigations have used simulation studies to investigate the ability of SNPs to detect association. Our own group has carried out two such studies to characterise the ability of marker polymorphisms to detect association with nearby variants influencing susceptibility to a disease (North et al. 2004; North et al. 2006). A previous report assessed the relative power of Affymetrix 100K, Affymetrix 500K and Illumina 300K chipsets to detect association (Pe'er et al. 2006). Since then, the Illumina 550K chipset has become available. A more recent report (Zaitlen et al. 2007) proposed a new method for multimarker analysis called WHAP and provided software to implement it on the Affymetrix 500K and Illumina 550K chipsets. It reported that, with a sample size of 2000, a disease prevalence of 0.01 and a relative risk of 1.5, the power of single SNP analysis to detect association at $p < 0.01$ with these chipsets was 0.92 and 0.98 respectively and the power using the WHAP method was 0.96 and 0.99. A second recent report (Magi et al. 2007) compared the extent to which the four chipsets are able to tag other SNPs in different ethnic groups by measuring the percentage of SNPs covered with $r^2 > 0.8$ and by measuring the mean r^2 between each SNP and its best tagging SNP. It was found that in a European population the Illumina 550K set provided 86% coverage, followed by the Illumina 300K (76%), Affymetrix 500K (64%) and finally Affymetrix 100K (32%).

*Corresponding author: Prof David Curtis, Adult Psychiatry, Royal London Hospital, Whitechapel, London E1 1BB, UK. Tel: +44 20 7377 7729. Fax: +44 20 7377 7316. E-mail: david.curtis@qmul.ac.uk.

Because microsatellites have a larger number of alleles than SNPs one might expect from a theoretical point of view that they could be more powerful to detect association (Sham et al. 2000). Against this, their higher mutation rate might make LD decay more swiftly. Nevertheless, it has been shown that LD can be detectable between microsatellites over long distances (Sherrington et al. 1991), implying that LD with a pathogenic variant might also be detectable in association studies of disease. Hence it is difficult to judge *a priori* the relative benefits of each type of marker. One would also want to consider issues such as the cost, speed and convenience of genotyping, factors which constantly change as new technology develops. If it were to be shown that one or other type of marker were better able to detect association then this in itself would act as a spur to further technological refinements. Although the current fashion seems to be to move towards SNPs in order to carry out systematic testing for association over a region and indeed over the whole genome, it might be that microsatellites could be useful in this role. The relative usefulness of SNPs and microsatellites has not previously been systematically studied.

Rather than reporting the proportion of polymorphisms tagged, the present study uses simulation methods to provide a real world assessment of the relative power of the Affymetrix 500K, Illumina 300K and Illumina 550K chipsets to detect an associated disease locus. Separately, it measures the extent of LD between microsatellites and nearby SNPs with a view to assessing the ability of microsatellites to detect association.

Methods

We have previously described in more detail the general approach of using simulations to assess the power of SNPs to detect association (North et al. 2006). In essence, the method consists of using real SNP genotypes from the HapMap project (HapMap Consortium 2003) to produce simulated data for a case-control study based on observed SNP allele frequencies and LD relationships. The scenario envisaged is that one SNP affects susceptibility to disease but has not been genotyped. Available to the investigator are the phase-unknown genotypes of nearby markers and the aim is to detect association with the disease phenotype in the context of a case-control association study.

Original SNP Genotypes

SNPs covering regions extending from 5 cM to 10 cM of chromosomes 1 to 5 were downloaded from the HapMap site (www.HapMap.org). Genotypes were taken for 60 unrelated subjects, who are parents in the 30 trios comprising the CEPH dataset. 1000 consecutive SNPs having minor allele frequency (MAF) of at least 0.05 were selected from the midst of each of these 5 regions. The cut-off of using MAF ≥ 0.05 was designed to exclude HapMap SNPs which were monomorphic or nearly monomorphic in this population. These 1000 SNPs were used as simulated disease loci and were studied along with the surrounding SNPs.

The HapMap SNPs were downloaded from the HapMap website (www.HapMap.org).

Estimation of Disease-Marker Haplotypes

For the 5000 SNPs to be studied we selected each SNP in turn to be treated as if it were a disease susceptibility locus. We wished to investigate the ability of other SNPs nearby to detect association with the disease phenotype. In order to gain an idea of the probability for at least one marker or group of markers to be able to detect the effect of the disease locus we carried out analyses with several neighbouring markers. For each selected disease locus we used neighbouring SNPs to act as markers and used marker sets of different sizes ranging from 1 to 4 contiguous markers to form a sliding window, while skipping the disease locus itself. We chose 6 sets of each size, 3 on either side of the disease locus, with the first marker shifting by one position each time. Thus, for single marker analyses we used 1 of 3 single markers on either side of the disease locus whereas for the 4-marker analyses we used 6 overlapping windows of 4 markers at a time. For each combination of disease and marker loci we obtained estimated haplotype frequencies in the downloaded genotypes using the SNPHAP program (www.gene.cimr.cam.ac.uk/clayton/software/). These haplotype frequencies estimated from real data were then used to generate simulated sets of marker genotypes, such as might be observed in case-control studies were the marker(s) to be typed in a sample in which the disease locus exerted an effect on susceptibility. No minimum frequency was set for these haplotypes, though obviously haplotypes with very low frequency would only rarely be sampled in the simulations.

Simulation of Genotypes

For each disease locus a penetrance of 0.01 was used for subjects having no copies of the disease allele, while for subjects having one or two copies of the disease allele the penetrance was set to 0.02. The allele frequencies of the disease locus were taken to be the observed frequencies of the SNP under consideration. As described previously (North et al. 2006), the expected proportions of cases and controls having 0, 1 or 2 copies of the disease allele were calculated using this transmission model under the assumption of Hardy-Weinberg equilbrium within the population. A simulated sample of 400 cases and 400 controls was then generated. Each case or control was first randomly allocated a number of disease alleles according to the probabilities equal to the expected proportions. Then two haplotypes bearing this number of disease alleles were sampled at random according to the estimated haplotype frequencies obtained from the SNPHAP program.

Use of Different Marker Sets

The procedure described was carried out assuming that 4 different samples of SNPs were available to use as markers. Firstly, we utilised all available HapMap SNPs. Alternatively, we restricted our use of markers to those included in different chipsets. The list of SNPs for the Affymetrix 500K array chipset was downloaded from the Affymetrix website (www.affymetrix.com) and the lists of the Illumina 300K and 550K chipsets were downloaded from the Illumina website (www.illumina.com). For each disease locus we obtained simulated marker genotypes using one of these four marker samples, i.e. all HapMap markers, Affymetrix 500K, Illumina 300K or Illumina 550K. For all four marker samples we used flanking SNPs as markers but did not include the actual disease locus genotypes in the data available to be analysed.

Analysis of Simulated Genotypes

The genotypes obtained from the above procedure were analysed using two methods: a test for heterogeneity of haplotype frequencies using RUNGC and logistic regression. For heterogeneity testing the GENECOUNTING program and RUNGC support program were used (Zhao et al. 2002; Curtis et al. 2006). The GENECOUNTING program estimates maximum likelihood haplotype frequencies from unphased multilocus genotypes, which may include multiallele genotypes and missing data. It also outputs a log likelihood for the dataset assuming these frequencies. The RUNGC program constructs a test for heterogeneity of haplotype frequencies by obtaining these maximised log likelihoods for the controls, the cases and the combined dataset. A likelihood ratio statistic (LRS) is derived as $2(L_{CASE} + L_{CONTROL} - L_{COMBINED})$ and this is taken as a chi-squared statistic with degrees of freedom equal to the difference in the number of haplotypes estimated to have non-zero frequency in the cases plus the controls compared to the combined dataset.

In addition to heterogeneity testing, logistic regression analysis was used to test for the main effects of each marker locus. The A allele at each marker SNP was arbitrarily chosen as a risk factor which might influence risk so that genotypes AA, AB and BB would correspond to exposure of 2, 1 or 0. No interaction terms of independent variables were included in this analysis. This method had been implemented within the simulation program and logistic regression was carried out to estimate the log likelihoods for the dataset assuming no genetic effect on risk or assuming that the risk allele at each marker locus exerted an independent effect, producing a LRS having degrees of freedom equal to the number of marker SNPs included in the analysis. For single markers, this method is asymptotically equivalent to the Armitage test for trend.

For each disease locus we analysed the 6 simulated sets of marker genotypes using both heterogeneity testing and logistic regression and then we noted whether any of the sets produced evidence for association at $p < 0.001$. This level of significance was chosen arbitrarily as being one which some researchers would regard as being of interest in terms of possibly indicating association for a plausible candidate gene and as being achievable in a reasonable proportion of trials with the chosen sample size and genetic model. We carried out this process for sets of 1, 2, 3 and 4 markers.

Obtaining Microsatellite Genotypes

The microsatellite genotypes were obtained from the CEPH genotype database (ftp://ftp.cephb.fr/ceph_genotype_db/ceph_db/Ver_10/) through a process which proved surprisingly complex. The *mkr* directory contains files providing information on all of the 32356 available markers. Markers whose names were D numbers were selected from the *mkr* files and those with more than two alleles and heterozygosity rate greater than 0.5 were selected as possibly being microsatellites. The positions of these microsatellites were then obtained from the UCSC Genome Bioinformatics table browser (http://genome.ucsc.edu/cgi-bin/hgTables), which provides an alias file and a map file. For a given region, the alias file gives aliases for the D numbers, which can then be found in the map file, which gives the microsatellite positions. The *asc* directory contains files which give genotypes for the subjects in the CEPH families. Individuals are identified by their family ID and their within-family ID, which differ from the ID used by HapMap. Pedigree information on the 30 CEPH trios included in the HapMap project is available at http://www.HapMap.org/downloads/samples_individuals/. This enables an individual with a HapMap subject id to be located within the CEPH genotype (*asc*) files and enabled us to obtain genotypes for putative microsatellites matched to the HapMap subjects for which we already had SNP data.

Assessing Ability of Microsatellites to Detect Association with Nearby SNPs

Our original intention had been to treat the microsatellites in the same way as we had treated marker SNPs and to produce simulated microsatellite genotypes based on a simulated disease locus effect of a test SNP and estimated LD relationships between that SNP and microsatellites nearby. However we were unable to proceed with this for two reasons. Firstly, it emerged that of the microsatellites typed in CEPH pedigrees very few had genotypes available for the parents whose SNP genotypes were available from HapMap. Other members of the pedigrees tended to be genotyped instead. This meant that there were not microsatellite genotypes available in the regions which we had studied for the SNP simulations. Secondly, we realised that with the relatively small number of subjects we had available estimates of SNP-microsatellite haplotypes might not be sufficiently accurate to produce meaningful simulations. If a microsatellite had a large number of alleles then the maximum likelihood estimate for the haplotype frequencies might be critically dependent on observations in just one or two individuals. If estimated haplotype frequencies are taken to represent the true population frequencies and are then used in a simulation procedure such as that implemented above then when a large sample is generated there is a risk of strongly exaggerating the extent of LD between the test SNP and marker microsatellites. This process can lead to the production of very unrealistic simulation results. Reports elsewhere

Table 1 D-numbers, start positions, marker type and heterozygosity rate are shown for the ten chromosome 1 microsatellites used for tests for LD with nearby SNPs.

D-number	Start position	Enzyme/ Marker type	Heterozygosity rate
D1S1597	13656694	(GATA)n	0.78
D1S164	33652218	(AC)n	0.78
D1S168	39762119	(AC)n	0.55
D1S162	50669089	(AC)n	0.84
D1S159	69991533	(AC)n	0.63
D1S1679	160628387	GGAA5F09/pcr	0.83
D1S1677	161826323	GGAA22G10/pcr	0.74
D1S178	230426715	(AC)n	0.64
D1S163	232925994	(AC)n	0.73
D1S180	239431757	(AC)n	0.88

highlight similar dangers when estimated haplotype frequencies are treated as if they accurately represented population frequencies (Curtis & Sham 2006; Curtis & Xu 2007).

Instead we adopted a different approach. Firstly, we identified all microsatellites on chromosome 1 which had been genotyped in at least 40 of the CEPH parents. Out of a total of 5,000 on chromosome 1 there were 10 which met this criterion and they are listed in Table 1. Next, we downloaded from HapMap all SNPs within 2 Mb of each microsatellite.

In order to measure the extent of LD between each microsatellite and a nearby SNP we calculated the significance of a likelihood ratio test for LD as implemented in the LDPAIRS program (Curtis et al. 2006). This calculates the log likelihood for genotypes at two loci assuming that there is no LD between them and then calculates the log likelihood assuming maximum likelihood estimates for the haplotype frequencies allowing for LD. Twice the difference between the log likelihoods is taken to be distributed as chi-squared with degrees of freedom equal to NM-N-M+1, where N and M are the numbers of alleles at each locus. We took $-\log(p)$ as providing some kind of indication as to whether or not there was evidence for some level of detectable LD between the microsatellite and the SNP and for each SNP we graphed this against the distance away from the microsatellite. Using this approach, a relatively small sample is not expected to produce artefactual evidence of LD as it would if simulation procedures were applied. The LDPAIRS program also outputs two measures of LD, Cramer's V and D' between the commonest alleles at the two loci. However for the present application we thought that the significance of the test for LD would provide a better guide to the likely ability of a microsatellite to demonstrate significant evidence for association in a case-control study in which the SNP exerted an effect on risk.

Results

Power of SNPs to Detect Association

Results from heterogeneity testing of haplotype frequencies and of logistic regression analysis were similar but the latter method was slightly more powerful. Hence only results from logistic regression analysis will be considered further. Likewise, results for the five different chromosomal regions studied were similar to each other and hence were pooled. Table 2 shows the proportion of the 5,000 simulated disease loci for which at least one of the six flanking marker sets tested was significant at $p < 0.001$. The following points can be made. Firstly, the overall power to detect association by this criterion is in the region of 0.4. Of course, this figure would be critically dependent on disease model and sample size and is not the main emphasis of this investigation. Secondly, we note that the power increases somewhat (by around 0.05) when we move from single marker analysis to two marker analysis. However using three or four markers does not produce any further notable increase in power to detect association. Finally, we note that the Affymetrix 500K chipset is somewhat less powerful (by around 0.07) than using all HapMap markers, that the Illumina 300K chipset loses a little less power (around 0.05) and that when using two marker analysis the Illumina 550K chipset loses minimal power (around 0.02).

Table 2 The proportions of 5000 simulated disease loci for which at least one of the six sliding window logistic regression analyses gave a p value of less than 0.001.

	Number of markers in sliding window			
SNP set	1	2	3	4
All HapMap	0.44	0.47	0.47	0.46
Affymetrix 500K	0.36	0.40	0.41	0.41
Illumina 300k	0.37	0.43	0.44	0.44
Illumina 550k	0.39	0.45	0.45	0.46

Extent of LD Between Microsatellites and Nearby SNPs

Figure 1 shows the extent of LD between the ten microsatellites investigated and the SNPs surrounding them, measured as $-\log(p)$. Even though only ten have been studied, it is clear that the patterns of LD vary widely between the microsatellites. For D1S159 and D1S1679 the pattern is perhaps close to what might be expected. There are a few SNPs which are very close (within a few kb) which show very strong evidence for LD whereas SNPs further away conform to chance expectation. By contrast, D1S162 appears to show evidence for LD with markers ranging up to 1.3 Mb away on one side but not with any markers on the other side of it. Other microsatellites, such as D1S1597, D1S164 and D1S168 show more moderate evidence for LD with nearby markers which arguably falls

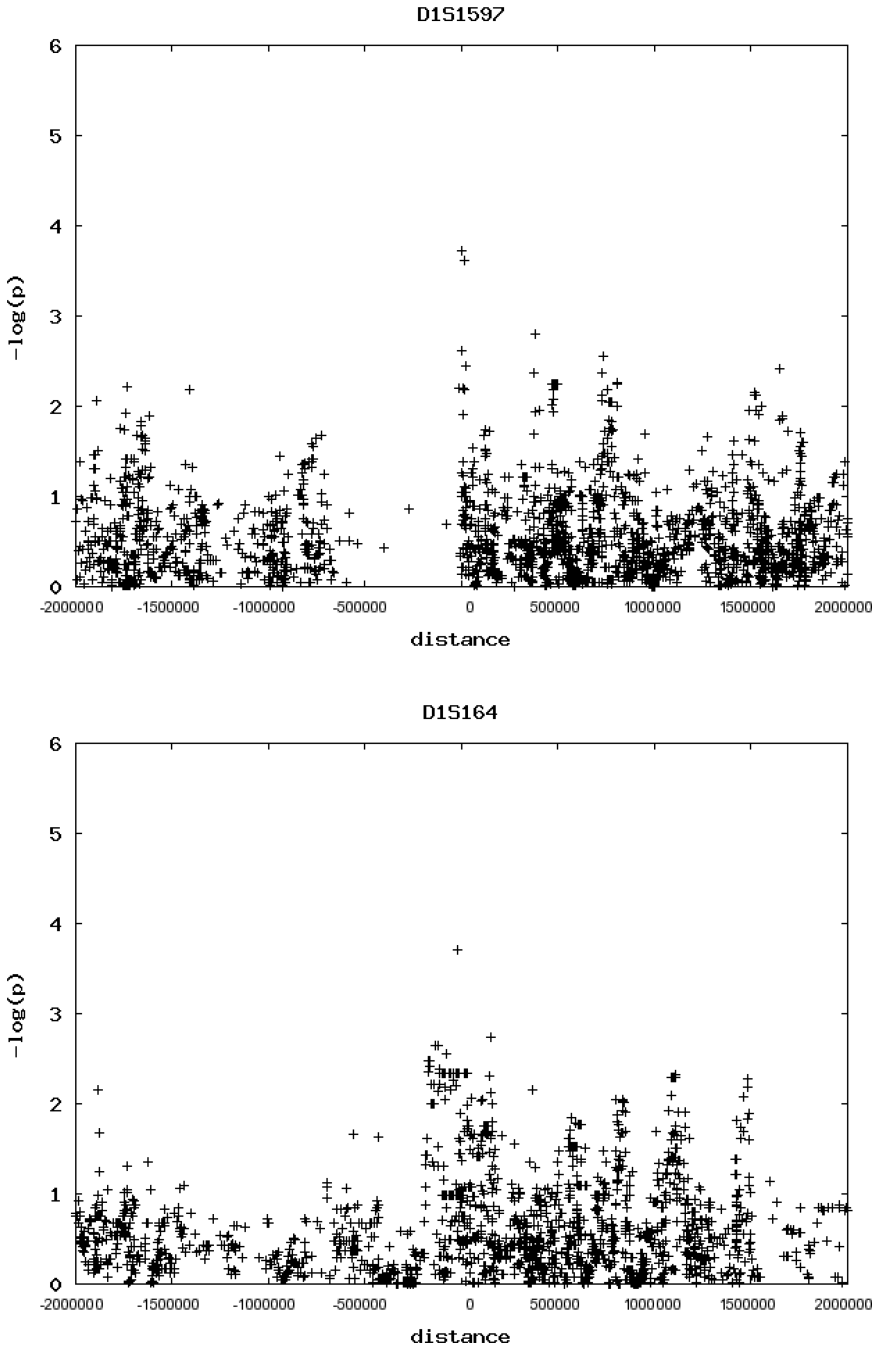

Figure 1 Values for $-\log(p)$ for the test for LD between microsatellites and SNPs lying within 2 Mb.

off more gradually with markers which are further away. Overall, we would argue that these results demonstrate two features. Firstly, only a small proportion of SNPs demonstrate LD with microsatellites even within a distance as small as 100 kb. Secondly, the pattern of LD relationships between microsatellites and SNPs is very inconsistent and in particular some microsatellites show quite strong evidence for LD with SNPs which are quite distant while others do not.

Discussion

The emphasis of this investigation is to assess the relative rather than absolute power of different marker sets to detect

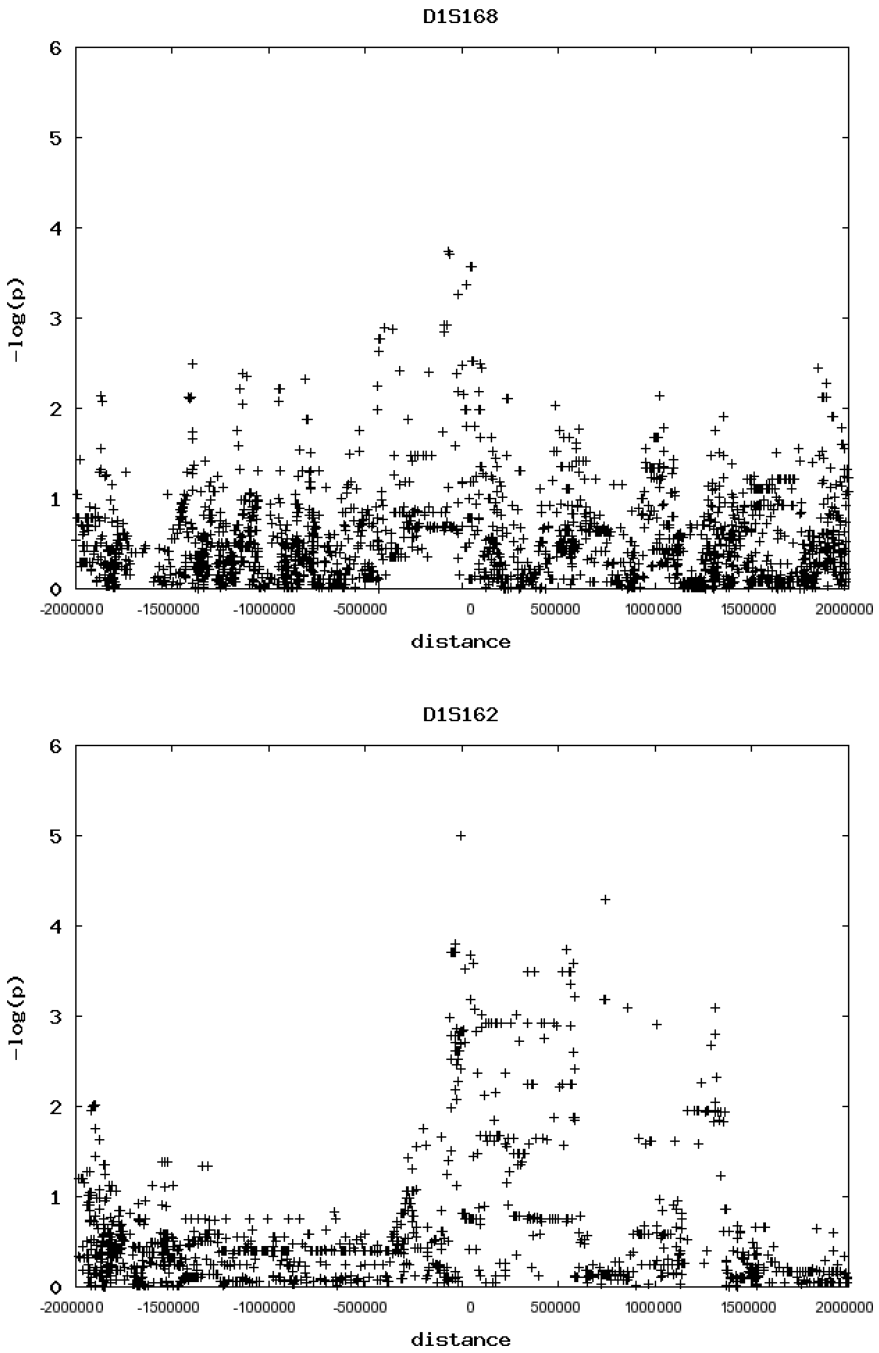

Figure 1 Continued

association of a marker with a disease phenotype. Nevertheless perhaps it is worth commenting that with the disease model and sample size used less than half of simulated loci produce evidence of association at the level of $p < 0.001$ with any nearby marker SNP. Where there is some degree of LD between the disease locus and marker one would expect the evidence for association to increase with increased sample size and/or increased genetic effect. However if there are susceptibility polymorphisms which have no or very weak LD with neighbouring polymorphisms then they will not be detected by association studies unless they themselves are genotyped.

Because we were interested in the relative power of the different marker chipsets, we have not investigated using other disease models or other methods of analysis. We might expect that while a different disease model could

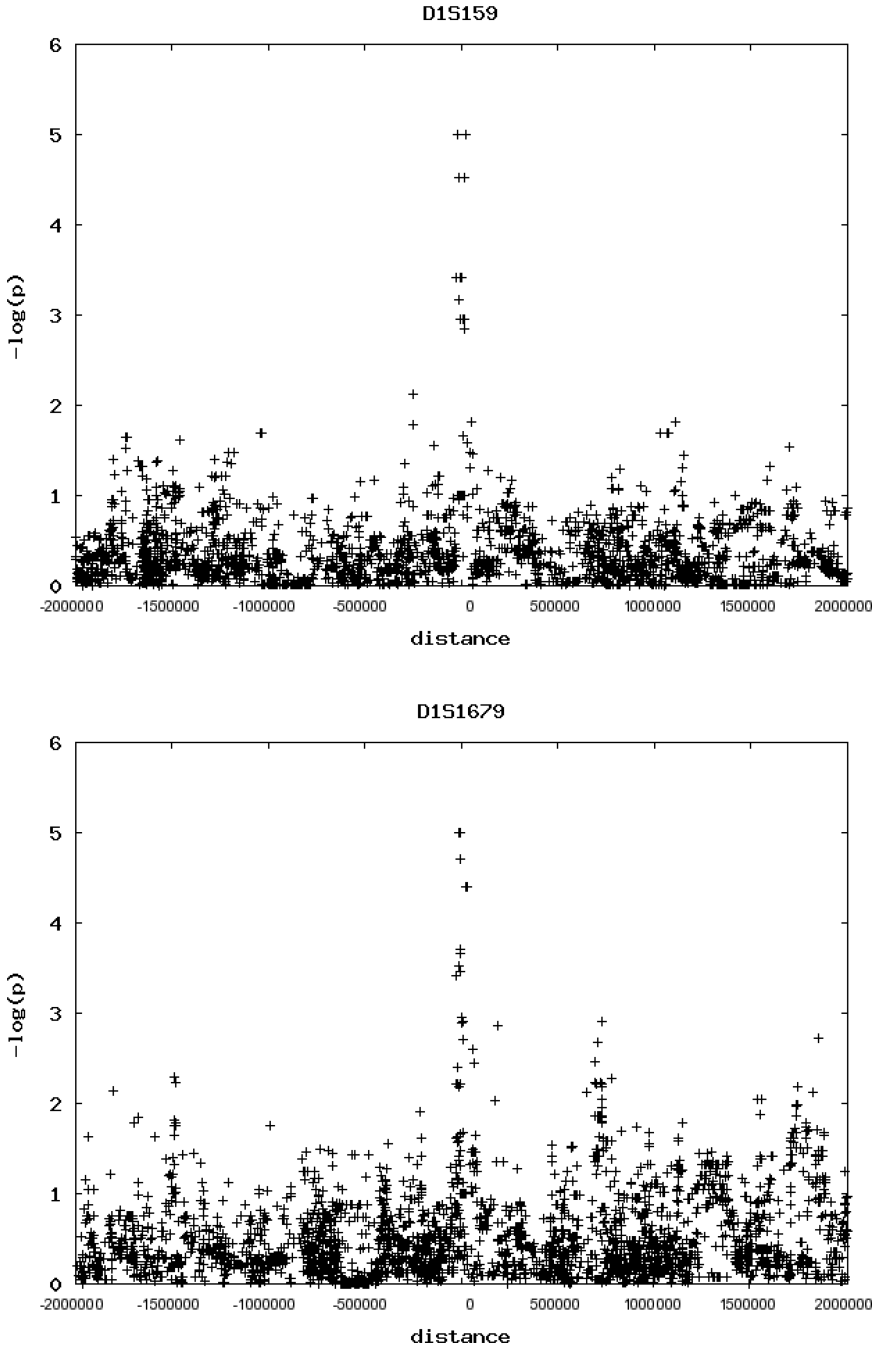

Figure 1 Continued

have a profound effect on the absolute power to detect association the relative power might not be greatly affected. Of course, we cannot be sure that there would not be circumstances in which some disease models might be relatively better suited to different marker chipsets and if this question were regarded as being of crucial importance then further studies could be performed. We note that the haplotype analysis using the asymptotic chi-squared distribution was less powerful than logistic regression analysis. It is to be expected that if no lower limit is set to estimated haplotype frequencies, as in the present case, then there may be an increased probability of both Type 1 and Type 2 errors when the asymptotic distribution is used and it has been reported elsewhere (Curtis, 2007) that the power of

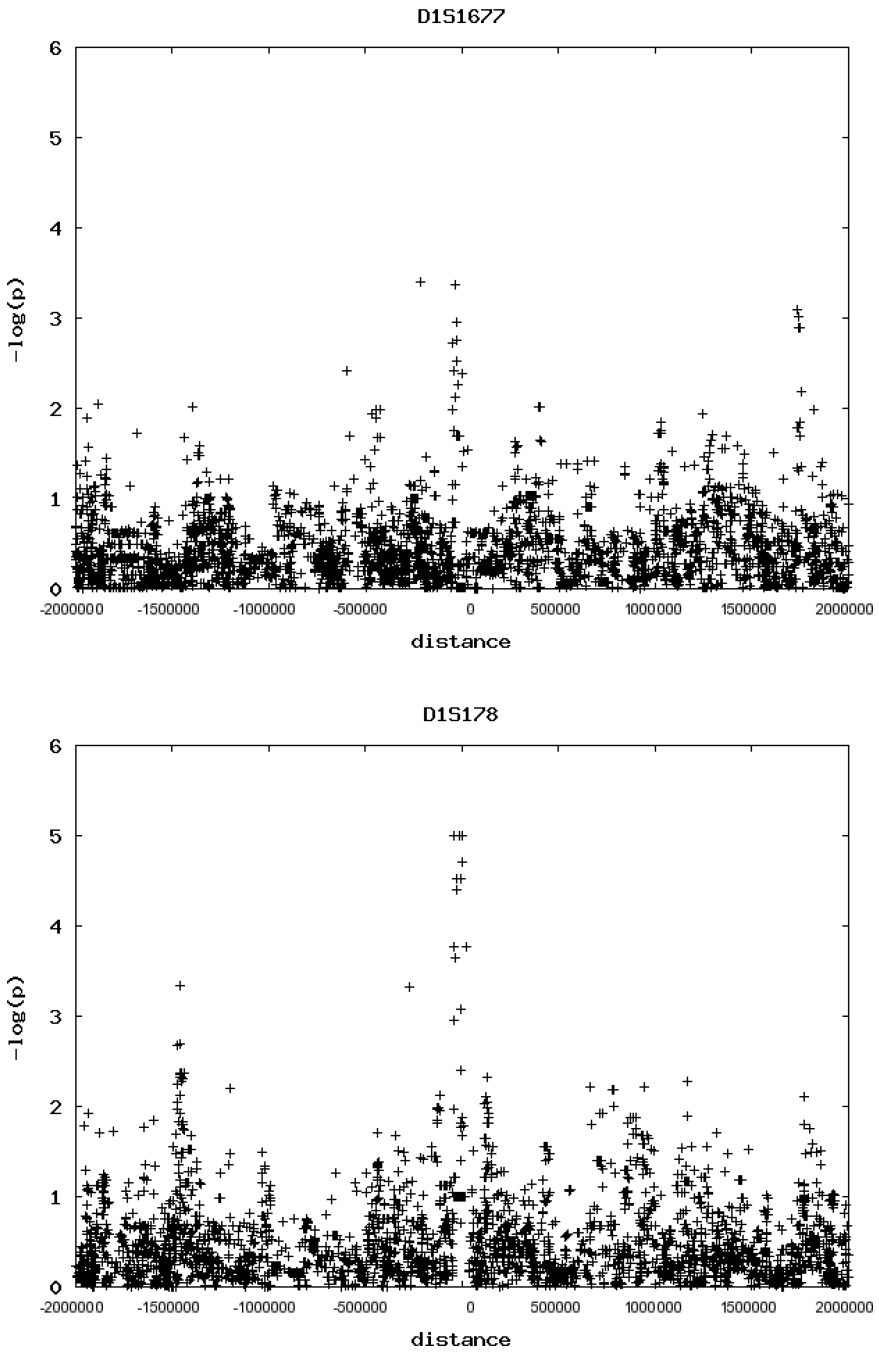

Figure 1 Continued

haplotype analysis is improved when permutation testing is used instead. We have not investigated the ability of newer multimarker methods to detect association (Minichiello & Durbin, 2006; Curtis, 2007; Zaitlen et al. 2007). These methods have yet to be taken up widely. Once again, it is not impossible that different marker chipsets could be better suited to different methods and if this were thought to be the case then further investigations could be carried out.

In terms of relative power, we conclude that all three of the chipsets studied have fairly good power relative to typing all available HapMap markers. In particular, the Illumina 550K chipset loses only minimal power, especially if two marker rather than single marker analysis is used. As

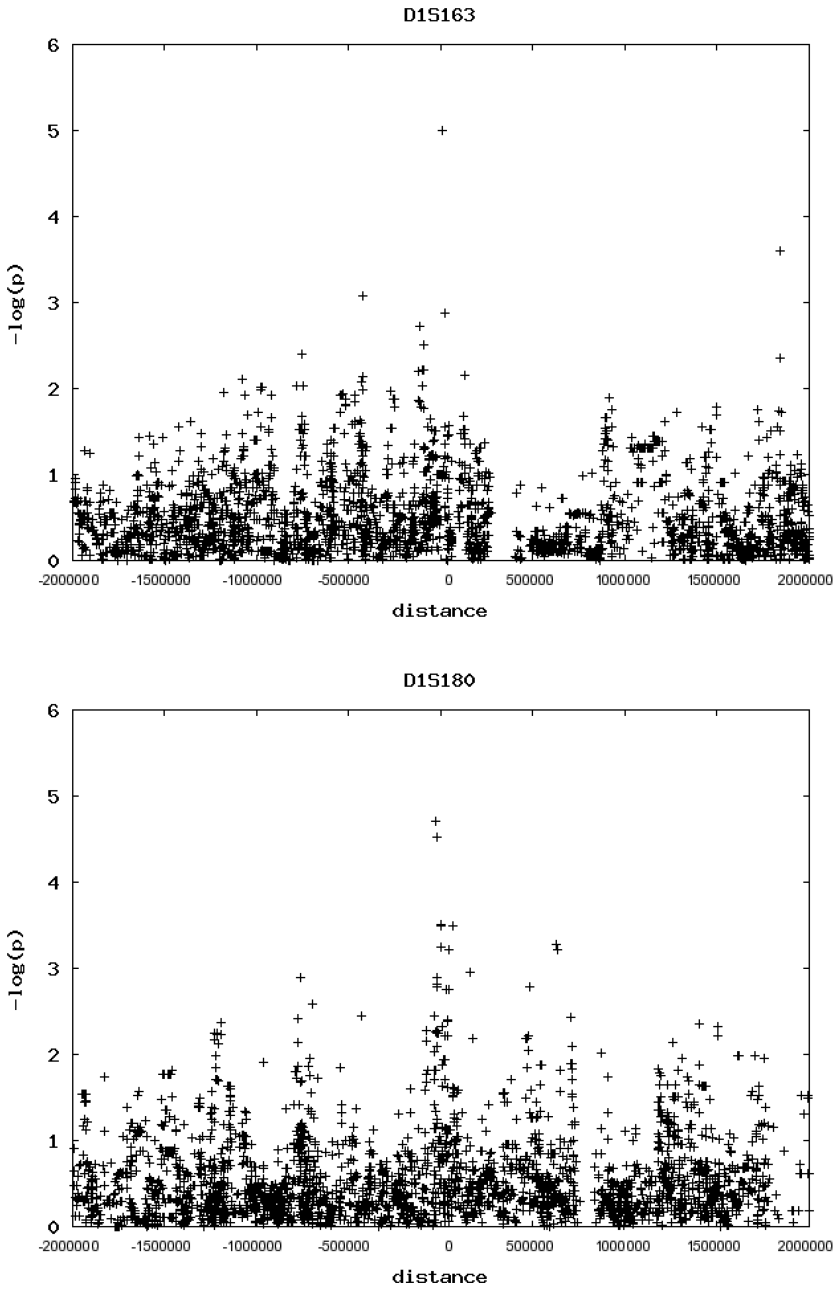

Figure 1 Continued

chipsets are introduced which include even more markers it may be useful to investigate how much additional power, if any, they yield.

The Illumina 550K markers have been specially selected to tag other SNPs which can be predicted from haplotypes of pairs of the Illumina markers. We deliberately chose to use an information-free method for combining information from multiple markers. Our reasoning was that, although the tagging method might allow the detection of association with known polymorphisms, we were interested in the ability of the markers to detect association with polymorphisms which might as yet be unknown. Even without introducing prior information about known LD relationships between pairs of Illumina markers and other HapMap SNPs we found that standard two locus logistic regression analysis using the Illumina 550K chipset yielded very good power.

With regard to the microsatellites, our conclusion is that they are not suitable for systematic tests for association. Each microsatellite demonstrates LD with only a small

proportion of SNPs lying within 100 kb. Thus if one of these SNPs had an effect on risk of affection there would only be a small probability that the microsatellite would demonstrate association in a case-control study. Since microsatellites on the CEPH database are spaced at an average of around 500 kb it seems that even if all known microsatellites were genotyped they would still be unable to detect association with the vast majority of known polymorphisms. An additional, more minor, problem is that they appear to demonstrate quite variable patterns of LD and would sometimes detect association with markers which were a large distance (> 1 Mb) away. This would make it more difficult to incorporate results from microsatellite genotyping into fine-mapping investigations. Although SNPs may also demonstrate variable patterns of LD and may sometimes demonstrate LD over long distances, we would argue that the fact that SNPs are far more numerous than microsatellites makes the utilisation of SNPs in association studies relatively less difficult. Where one is interested in a particular candidate gene or region it may be worth typing microsatellites nearby in case they happen to enhance evidence for association. However in general we conclude that the use of microsatellites for systematic association studies is problematic.

Acknowledgements

AV was supported by Wellcome Trust Project Grant, Grant No. 076392. JK was supported by an MRC Bioinformatics Training Fellowship, Grant No. G0501329.

References

Curtis, D. (2007). Comparison of artificial neural network analysis with other multimarker methods for detecting genetic association. *BMC Genet* **8**(1), 49.

Curtis, D., Knight, J. et al. (2006). Program report: GENECOUNTING support programs. *Ann Hum Genet* **70**(Pt 2), 277–9.

Curtis, D. & Sham, P. C. (2006). Estimated haplotype counts from case-control samples cannot be treated as observed counts. *Am J Hum Genet* **78**(4), 729–30; author reply 728–9.

Curtis, D. & Xu, K. (2007). Minor differences in haplotype frequency estimates can produce very large differences in heterogeneity test statistics. *BMC Genet* **8**, 38.

HapMap Consortium (2003). The international HapMap project. *Nature* **426**, 789–796.

Magi, R., Pfeufer, A. et al. (2007). Evaluating the performance of commercial whole-genome marker sets for capturing common genetic variation. *BMC Genomics* **8**, 159.

Minichiello, M. J. & Durbin, R. (2006). Mapping trait loci by use of inferred ancestral recombination graphs. *Am J Hum Genet* **79**(5), 910–22.

North, B. V., Curtis, D. et al. (2004). Further investigation of linkage disequilibrium SNPs and their ability to identify associated susceptibility loci. *Ann Hum Genet* **68**(Pt 3), 240–8.

North, B. V., Sham, P. C. et al. (2006). Investigation of the ability of haplotype association and logistic regression to identify associated susceptibility loci. *Ann Hum Genet* **70**, 893–906.

Pe'er, I., de Bakker, P. I. et al. (2006). Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet* **38**(6), 663–7.

Sham, P. C., Zhao, J. H. et al. (2000). The effect of marker characteristics on the power to detect linkage disequilibrium due to single or multiple ancestral mutations. *Ann Hum Genet* **64**(Pt 2), 161–9.

Sherrington, R., Melmer, G. et al. (1991). Linkage disequilibrium between two highly polymorphic microsatellites. *Am J Hum Genet* **49**(5), 966–71.

Zaitlen, N., Kang, H. M. et al. (2007). Leveraging the HapMap correlation structure in association studies. *Am J Hum Genet* **80**(4), 683–91.

Zhao, J. H., Lissarrague, S. et al. (2002). GENECOUNTING: haplotype analysis with missing genotypes. *Bioinformatics* **18**(12), 1694–5.

Received: 18 September 2007
Accepted: 18 January 2008