Genetic Epidemiology 31: 179–188 (2007)

Effects of Single SNPs, Haplotypes, and Whole-Genome LD Maps on Accuracy of Association Mapping

Nikolas Maniatis,* Andrew Collins, and Newton E Morton

Human Genetics Division, University of Southampton, Southampton General Hospital, Southampton, UK

We describe an association mapping approach that utilizes linkage disequilibrium (LD) maps in LD units (LDU). This method uses composite likelihood to combine information from all single marker tests, and applies a model with a parameter for the location of the causal polymorphism. Previous analyses of the poor drug metabolizer phenotype provided evidence of the substantial utility of LDU maps for disease gene association mapping. Using LDU locations for the 27 single nucleotide polymorphisms (SNPs) flanking the CYP2D6 gene on chromosome 22, the most common functional polymorphism within the gene was located at 15 kb from its true location. Here, we examine the performance of this mapping approach by exploiting the high-density LDU map constructed from the HapMap data. Expressing the locations of the 27 SNPs in LDU from the HapMap LDU map, analysis yielded an estimated location that is only 0.3 kb away from the CYP2D6 gene. This supports the use of the high marker density HapMap-derived LDU map for association mapping even though it is derived from a much smaller number of individuals compared to the CYP2D6 sample. We also examine the performance of 2-SNP haplotypes. Using the same modelling procedures and composite likelihood as for single SNPs, the haplotype data provided much poorer localization compared to single SNP analysis. Haplotypes generate more autocorrelation through multiple inclusions of the same SNPs, which could inflate significance in association studies. The results of the present study demonstrate the great potential of the genome HapMap LDU maps for high-resolution mapping of complex phenotypes. *Genet. Epidemiol.* 31:179–188, 2007.

Key words: *CYP2D6*; association mapping; linkage disequilibrium map; LDU; SNPs; haplotypes

Abbreviations used: df, degrees of freedom; LD, linkage disequilibrium; LDU, LD unit; SNP, single nucleotide polymorphism

Contract grant sponsor: US National Institute of Health.

*Correspondence to: Nikolas Maniatis, Human Genetics Division, Southampton General Hospital, University of Southampton, School of Medicine, Duthie Building (MP808), Southampton SO 16 6YD. E-mail: N.Maniatis@soton.ac.uk

Received 20 March 2006; Revised 25 August 2006; Accepted 31 October 2006
Published online 6 February 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20199

INTRODUCTION

Over the last few years, association mapping of disease genes has developed into one of the most dynamic research areas of human genetics. The central aim is to identify genes which contribute to complex diseases. The first step identifies candidate regions in the genome that are associated with the disease of interest. Subsequently, association mapping focuses on finer localization of disease determinants and the ultimate identification of the causal variants. In contrast to linkage analyses, which permit comparatively low-resolution mapping with the available family resources, efforts to map genes of complex diseases are concerned with exploiting linkage disequilibrium (LD) between markers and putative disease-predisposing loci, usually from population samples. LD analysis offers the prospect of fine scale localization of genetic polymorphisms of medical importance, particularly when single nucleotide polymorphisms (SNPs) are densely typed in a candidate region.

The development of LD maps led to the characterization of the LD structure by assigning a linkage disequilibrium unit (LDU) location for each marker SNP [Maniatis et al., 2002]. These maps are analogous to linkage maps and have distances which increase monotonically with physical maps but are superior in representing the pattern of LD rather than just recombination [Lonjou et al., 2003; Tapper et al., 2005]. The properties of these maps were first examined by Zhang et al. [2002], who found a remarkable agreement between LDU steps and sites of meiotic recombination using data of Jeffreys et al. [2001],

which confirmed the location of recombination hotspots by sperm typing. Subsequently, an association mapping approach that utilizes LDU maps was developed as a composite likelihood approach which models association across many markers [Maniatis et al., 2005]. The utility of LDU maps was first examined by simulating each SNP as causal from two existing real SNP data sets [Maniatis et al., 2004]. It was shown that greater power is achieved when mapping within an LDU map compared to a map in kb, especially in a densely typed region that is characterized by intense recombination hotspots. Several factors determine the power to identify a candidate region for a gene contributing to a particular phenotype, and within that region to localize a causal polymorphism. The use of an appropriate metric is essential. Maniatis et al. [2005] demonstrated the superiority of the association z over regression and correlation, for a 2×2 table of affection status by allelic dichotomy. The utility of this method was investigated by refining the localization of polymorphisms controlling the poor-metabolizer phenotype in the CYP2D6 gene. Previous studies using single markers and haplotypes identified a 390 kb region associated with this phenotype. As a proof of principle, 27 SNPs on chromosome 22, which cover an 880 kb region flanking the CYP2D6 gene with known location, were analysed. Using a metric LDU map, the most common functional polymorphism within the gene was estimated to be 15 kb from its true location, within a 95% confidence interval of 172 kb. Favourable results using alternative association mapping strategies have also been reported on the same data [Morris, 2005; Waldron et al., 2006]. Having investigated the properties of this method for association mapping using single SNPs, the present study evaluates the utility of haplotypes, which have received a great deal of attention over the last few years.

Data on about 0.6 and 1.2 million SNPs (release 13, 16, respectively) in four populations were publicly released by the HapMap project as part of their effort to foster the discovery of sequence variants that affect common diseases, facilitate development of diagnostic tools, and enhance our ability to choose targets for therapeutic intervention [International HapMap Consortium, 2003]. The enormous body of data created by the HapMap Project enables the creation of high-resolution, population-specific LD maps, and their locations in LDU can be directly used in association mapping studies to narrow a candidate region and increase the precision of localization. Therefore, the objective of this paper is twofold: firstly, to follow up the study by Maniatis et al. [2005] which was based on single SNP tests by investigating the extent to which haplotype data increase the power of LDU maps in the much-studied association mapping case of the CYP2D6 gene [Hosking et al., 2002]; secondly, we compare the performance of a high-density HapMap-derived LDU map with a sample-specific LDU map (CYP2D6 region). Such a comparison throws light on direct utilization of whole genome HapMap-derived LDU maps in association mapping of common diseases.

MATERIALS AND METHODS

DATA

The CYP2D6 gene affecting drug-metabolizing activity was introduced by Hosking et al. [2002] as a test of association mapping in a random sample of 1,018 Caucasians. None of the 27 tested SNPs are within the *CYP2D6* gene. From the 1,018 individuals, 41 are identified as slow metabolizers and therefore are called *affected*, and the remaining 977 are called *normal*. The *CYP2D6* locus on chromosome 22q13.1 metabolizes about 20% of commonly prescribed drugs [Evans et al., 2000]. There are four functional polymorphisms within the gene (G1846A, delA2548, delT1707 and A2935C), predicting 99% of slow metabolizers [Sachse et al., 1997]. In this study we consider the location of the most common functional polymorphism (G1846A, 20.7% allele frequency) as the true location of the gene. A full description of the data, including the *rs* numbers, is given elsewhere [Hosking et al., 2002].

LDU MAPS

The LD maps [Maniatis et al., 2002] assign markers to locations in LDU that describe the underlying structure of LD in the form of a metric map with additive distances. Therefore, every SNP in the data is assigned two locations, one in kb and the other in LDU. The construction of these maps is based on pairwise marker association and therefore any phenotypic information is completely ignored. The theory for constructing LD maps and the LDU locations of the 27 SNPs for the CYP2D6 region are given in Maniatis et al. [2005]. An LDU map of the whole chromosome 22 was

also constructed using public release 16 of the Phase I data from the HapMap Project. The HapMap data (http://www.hapmap.org/) were obtained on 60 parental DNA samples from Utah Mormons of north-western European ancestry collected by the Centre d'Etude du Polymorphisme Humain (CEPH). A total of 13,959 out of 19,017 SNP genotypes were used for LDU map construction after a screening procedure that rejected markers with $\chi_1^2 > 10$ for the Hardy-Weinberg test or a minor allele frequency less than 5%. The CEPH population was chosen because the 27 SNPs within an 880 kb region that flanks the CYP2D6 gene on human chromosome 22 were typed in 1,018 Caucasians. Therefore, two LDU maps were created based on pairwise marker-by-marker association, one from the CYP2D6 random sample described above and the other from 13,959 SNPs covering the entire chromosome 22 based on 60 parental individuals (HapMap). Throughout this study, these two maps are called CYP and HapMap, respectively. Map locations in kb of the most common functional polymorphism within the gene and the 27 predictive SNPs represent distances from SNP1 relative to the finished human genome sequence assembly (NCBI build 34, UCSC July 2003, http://genome.ucsc.edu/). Table I shows the kb and the HapMap LDU maps that were assigned for each of the 27 SNPs.

ASSOCIATION MAPPING

Single SNPs. Having ignored the phenotype and used the pairwise marker-by-marker association to create the CYP and HapMap LDU maps, we adapted this metric to compute an association metric $\hat{z}$ from the 2×2 table between the poor-metabolizer phenotype (0,1 in this case) and the two alleles of each marker SNP as: $\hat{z} = |D|/f(1-R)$, where D is the covariance between affection status and the marker alleles, f is the frequency of affected individuals and R is the minor allele frequency [Maniatis et al., 2005]. In this sample $f = 0.04 = 41/1018$, but this frequency may vary somewhat due to incomplete typing at a given marker. This method is based on single SNPs, and thus the number of tests performed is equal to the number of markers. For the ith SNP ($i = 1,\ldots,27$), the observed association $\hat{z}_i$ has an expectation z_i estimated by the Malecot model as: $z_i = (1-L)Me^{-\varepsilon d_i} + L$, where ε is the exponential decline of association with distance d_i in kb or LDU. The parameter M (intercept) reflects a monophyletic or polyphyletic origin of susceptibility alleles (i.e. proportion of disease alleles transmitted from founders): it is 1 if disease alleles are monophyletic or less than one if there are multiple mutations at the disease locus. The parameter L (asymptote) is the association at large distance. It reflects the bias due to definition of $\hat{z}$ as positive, which is necessarily >0. However, the object of this analysis is to estimate S, which is the location of the disease gene in the marker map. This parameter is introduced by substituting distance $d_i = \Delta(S_i – S)$, where S_i is the location of the ith marker and can be expressed either in kb or in LDU locations obtained from the CYP or HapMap LDU maps. The Kronecker Δ is used for map direction and assures a correct sign $\Delta = 1$ if $S_i \geq S$ or -1 if $S_i < S$. Therefore, the model becomes $z_i = (1-L)Me^{-\varepsilon\Delta(S_i - S)} + L$. The analysis of these data yielded $M \ll 1$ (0.82), reflecting the four functional polymorphisms that are involved in CYP2D6, which implies several ancestral alleles.

Given the observed associations $\hat{z}_i$ the Malecot parameters are estimated iteratively using composite likelihood which evades the heavy Bonferroni correction by combining information over all loci as $\Lambda = \Sigma K_i(\hat{z}_i - z_i)^2$, where $\hat{z}$ and z are the observed and expected association values, respectively, at the ith marker SNP. Their squared difference is weighted by an amount of information, K_i, which is estimated as: $K_z = \chi_1^2/\hat{z}^2$, where χ_1^2 is the Pearson's χ_1^2 from the 2×2 table (affection status by SNP alleles, Table I). Following Maniatis et al. [2005], we used four different sub-hypotheses of the Malecot model to test the existence of a causal polymorphism. The baseline is sub-hypothesis or Model A with none of the parameters estimated. The intercept has $M = 0$ and represents the null hypothesis of no association across the region. The parameter L is not iterated but can be predicted (L_p) from the mean deviation of information K_i. Model B conforms to Model A but with L iterated. It follows that any increase in L above the predicted asymptote L_p provides evidence of a causal polymorphism within the significant region in question, but without precise localization. Model C allows the estimation of both M and S. Model D is as Model C but with L iterated. Therefore, the A–B contrast tests for significance in the region, while the contrasts A–C and A–D test for a disease determinant at location S, or in the present study, the consensus location of CYP2D6. The location error in this case is the estimated distance from the model (S) to the

TABLE I. The kb and HapMap LDU maps of the CYP2D6 region

SNP[a]	kb	HapMap LDU	single SNPs χ^2_1	2-SNP haplotypes[b] χ^2_1
SNP1	0	0.000	5.5	—
SNP2	88	0.000	7.1	3.3
SNP3	177	0.000	2.0	2.6
SNP4	205	0.000	4.5	1.7
SNP5	238	0.376	1.1	2.8
SNP6	252	0.635	1.4	1.3
SNP7	252	0.635	0.8	0.7
SNP8	283	0.774	13.0	4.7
SNP9	294	1.049	54.1	28.3
SNP10	334	1.049	18.6	32.4
SNP11	368	1.456	32.1	46.9
SNP12	389	1.559	109.7	54.4
SNP13	423	1.609	42.5	54.6
SNP14	481	1.609	39.3	21.9
SNP15	500	1.615	41.7	113.7
SNP16	510	1.615	21.7	19.1
SNP17	510	1.615	35.9	16.5
CYP2D6	525			
SNP18	539	1.682	123.6	58.5
SNP19	557	1.689	26.2	100.3
SNP20	657	1.809	186.9	79.2
SNP21	675	1.843	13.1	72.0
SNP22	714	4.243	0.0	3.1
SNP23	740	4.782	0.2	0.2
SNP24	799	5.135	2.2	10.8
SNP25	823	5.303	0.5	0.1
SNP26	853	5.392	3.0	0.8
SNP27	879	7.718	1.2	1.2

[a]See Hosking et al. [2002] for database IDs and primers.
[b]Adjacent SNPs, e.g. SNP pairs 1 and 2, 2 and 3.
Pearson's χ^2_1 values from the 2×2 table between the PM phenotype and single marker alleles or haplotypes

location of the most common functional polymorphism (G1846A) at 525.3 kb.

The significance for the three contrasts is tested by the use of χ^2. For example the A–C test in large-sample theory has a $\chi^2_2 = (\Lambda_A - \Lambda_C)/V_C$, where V_C is the residual error variance of Model C and is computed by dividing the weighted sum of squares with the degrees of freedom m to give $V = \Lambda_C/m$. The degrees of freedom m equals the number of SNPs minus the number of parameters in the model k (e.g. model C has two parameters and thus the A–C test has a χ^2 with 2 df). For large samples, an F test $= \chi^2/k$. However, an F-test is more reliable than χ^2 when m is small. Here we only have 27 SNPs and hence 27 tests for the single SNP analysis. We recognize that estimates of χ^2 that were previously published by Maniatis et al. [2005] are based on a small number of marker tests and therefore we now implement an F test for computing significance. The F-value is estimated as the ratio of the between models mean square to the error mean square (error variance V). For the A–C contrast the significance test is: $F(k,m) = \frac{\Lambda_A - \Lambda_C}{k}/V_C$. The F tests for the A–B and A–D contrasts can be computed the same way using the corresponding values. Subsequently, we converted these F-values to a χ^2_1 (Appendix 1). This was done by obtaining the corresponding probability (P) of the F-test using a sub-routine from Press et al. [1994]. Then the χ^2_2 with 2 degrees of freedom is simply $-2\ln(P)$. The χ^2_2 is then converted to χ^2_1 using the Hastings approximation [Abramowitz and Stegun, 1964]. Therefore, the significance of the A–B, A–C and A–D contrasts is now tested by the use of χ^2_1 which is corrected for m. The 95% confidence interval (CI) for the estimated location $\hat{S}$ was obtained as: $\hat{S} \pm t$ SE, where t is the tabulated value of Student's-t test for m degrees of freedom and $P = 0.05$. The empirical standard error of parameter $\hat{S}$ is $\mathrm{SE} = \sigma_S\sqrt{V}$, where σ_S is the nominal standard error of $\hat{S}$ estimated by quadratic approximation of the composite likelihood. The corresponding lod for every point with S specified can also be estimated as: $\chi^2_1/2\ln 10$. The χ^2_1 is obtained from the $F(k,m) = [\Lambda_C(S) - \Lambda_C(\hat{S})]/V_C$, where $\Lambda_C(\hat{S})$ and V_C are the composite likelihood and error variance under model C when $\hat{S}$ was estimated. Both values are constant, while estimates of $\Lambda_C(S)$ were obtained by fitting the Malecot model to specified values of S in kb or LDU. Therefore, the lod surfaces for given values of S can be obtained for both maps and models. These surfaces can also be used to estimate the lod support interval as an alternative to the confidence interval. For comparison with the 95% CI the corresponding 95% support interval is the lod that equals 0.834, which is the lod for the 95% tabulated χ^2_1 (3.84/2ln10).

HAPLOTYPES

Haplotypes have received enormous attention over the last few years. Although the use of composite likelihood allows the analysis of single markers tests simultaneously, haplotypes raise important problems of definition and estimation when they are used in association mapping (e.g. combined information from multiple haplotypes). In this study, we have modelled the simplest case, considering only adjacent SNPs (2-SNP haplosets), e.g. SNP pairs 1,2 and 2,3. The midpoints of their

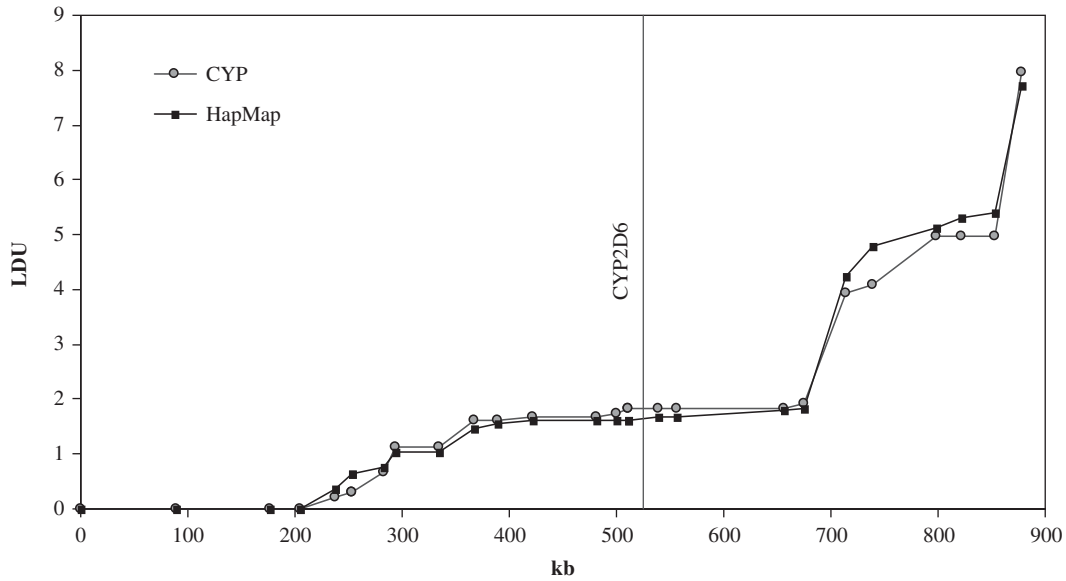

Fig. 1. The graph of the LDU map for the *CYP2D6* region. Vertical line indicates the location of the locus at 525.3 kb.

kb and LDU locations are assigned to each haploset. The statistical inference of haplotypes was conducted by Hill [1974] for a pair of diallelic loci in a panmictic population. Using this algorithm, the 3×3 table of genotypes for each pair of SNPs (AACC, AACc, AAcc), (AaCC, AaCc, Aacc), (aaCC, aaCc, aacc) was reduced to a 2×2 haplotype table (AC, Ac, aC, ac) for affected and normal individuals separately. Subsequently, we dichotomized the four haplotypes, AC, Ac, aC, ac, from each haploset by selecting the haplotype with the greatest value of Pearson's χ_1^2 by the z test. This test was based on the 2×2 table of affection status by haplotype dichotomy (Table I), for example, AC vs. Ac+aC+ac, taking the estimated haplotype frequencies from affected and normal as counts. Estimates of z for each of the 26 haplosets and their midpoint locations in kb and LDU were used under the same modelling procedures as for single SNPs. No correction was used for selecting the most significant haplotype, and so this procedure might be expected to exaggerate significance, as well as inflating the covariance between adjacent pairs that share the intervening marker.

RESULTS

The block-step structure of the *CYP2D6* region can be presented graphically by plotting both the HapMap and CYP LDU locations on the kb map (Fig. 1). There are 277 SNPs in the HapMap LDU map that cover the CYP2D6 region, but only the corresponding LDU locations for the 27 SNPs are plotted. Similar patterns were observed even though the samples that created the two maps were remarkably different in sample size and SNP density (277 vs. 27 SNPs and 60 vs. 1018 individuals). Most importantly, these two maps yielded the same LDU length, which demonstrates additivity of map distances and robustness to SNP density.

Table II shows the localization of CYP2D6 when SNP locations are expressed in kb and LDU. The A–C contrast shows a large increase in χ^2 when the data are fitted to the CYP LDU map ($\chi_1^2 = 74.1$), compared to the map in kb ($\chi_1^2 = 47.4$). The location error is only 14.9 kb, compared with an error of 57 kb in the kb map. The 95% interval is also considerably worse for kb, since the true location is not included within those limits. A smaller error variance and greater power ($\chi_1^2 = 76.4$) is observed with HapMap LDU locations compared to CYP LDU. The location ($\hat{S}$) was estimated to be 524.8 kb, which is on top of the CYP2D6 gene and less than 1 kb away from the location of the most common functional polymorphism (525.3 kb). This decrease in location error, from 14.9 to 0.5 kb, is the consequence of mapping within the LDU locations that were obtained from the HapMap data, instead of the locations that were based on marker-by-marker association of the CYP2D6 data. Although the

high resolution of the HapMap data greatly refined localization of the gene, it failed to reduce the confidence interval further. The 95% CI is 20 kb wider for the HapMap LDU (197 kb) compared to the CYP LDU map (177 kb). It is anticipated that new releases of HapMap will contain higher marker densities and provide finer LD structure which may reduce this interval even further. The computation of the 95% LOD support intervals yielded smaller difference (4 kb) between the two LDU maps compared to the CI (20 kb). However, these intervals were somewhat greater than the 95% CI (Table II). This is because the latter is computed using a normal theory approximation, while the LOD support tends to be more conservative, since SNPs within a block will have the same lod because they have identical locations in LDU. Therefore, computation of the support interval can reflect the block-step pattern of the region. The CYP and HapMap support intervals are very similar as a consequence of their similarity in LDU structure (Fig. 1). Plotting the lods against the LDU map may reveal different maxima other than the maximum likelihood estimation and that may have important applications to genome-wide scans where association mapping may give rise to several causal sites. In a simulation study, investigation of false-positive indications showed that the chi-square distribution yielded an acceptable goodness of fit, demonstrating that composite likelihood works very well despite its assumption of independence among the marker SNPs [Maniatis et al., 2004].

The same modelling procedures on the HapMap LDU map for associations between affection status and dichotomized 2-SNP haplotypes increased substantially the location error under model C (152 kb, Table III). Although the 95% confidence and support intervals were considerably smaller that the single SNP analysis, the limits for the CI did not include the CYP2D6 gene, while the support interval gave a limit only 6 kb (531 kb) away from the functional polymorphism (525.3 kb). Unlike the single SNP analysis, great inconsistencies between models C and D were found with the use of haplotypes. Model D yielded better localization with an error of 19 kb but considerably higher than the 0.3 kb error from single SNPs. The 95% CI and support intervals (162 and 163 kb, respectively) were also reduced compared to model C, but again their limits only just covered the gene. Table II shows that 2-SNP haplotypes did not perform as well as the single SNPs using the HapMap LDU map but the same holds true for the CYP LDU and map (result not shown). The analysis of the 2-SNP haplotypes yielded great inconsistencies and increased loca-

TABLE II. Localization of *CYP2D6* using different maps

Map	Contrast	V	χ^2_1	$\hat{S}_{kb}$	Location error $\lvert\hat{S} = 525.3\rvert_{kb}$	95% confidence interval in kb	95% LOD support internal in kb
kb[a]	A–C	3.16	47.4	468.3	57.0	427–510[b] (83)	427–521[b] (94)
CYP LDU	A–C	1.06	74.1	510.4	14.9	406–583 (177)	407–603 (196)
HapMap LDU	A–C	0.97	76.4	524.8	**0.5**	372–569 (197)	372–572 (200)

V is the error variance; χ^2_1 with 1 degree of freedom tests association at location $\hat{S}$ (see materials and methods for contrasting models A and C); 525.3 kb is the true location of the common polymorphism; values in parentheses are the total width for the confidence intervals.
[a]There is only one kb map, based on UCSC July 2003 assembly.
[b]Interval does not include the *CYP2D6* locus

TABLE III. Comparison of single SNPs and haplotypes

Analysis	Map	Contrast	V	χ^2_1	$\hat{S}_{kb}$	Location error $\lvert\hat{S} = 525.3\rvert_{kb}$	95% confidence interval in kb	95% LOD support internal in kb
Single SNPs	HapMap LDU	A–C	0.97	76.4	524.8	0.5	372–569 (197)	372–572 (200)
		A–D	0.99	71.1	525.0	0.3	372–570 (198)	372–601 (229)
Haplotypes	HapMap LDU	A–C	0.54	83.6	372.3	153.0	367–521[a] (152)	369–531 (162)
		A–D	0.43	82.9	506.3	19.0	372–534 (162)	371–534 (163)

V is the error variance; χ^2_1 with 1 degree of freedom tests association at location $\hat{S}$ under models C and D (see materials and methods for contrasting models); 525.3 kb is the true location of the common polymorphism; values in parenthesis are the total width for the confidence intervals.
[a]Interval does not include the *CYP2D6* locus.

tion errors compared to single SNPs. These differences were also accompanied with increased estimates of χ_1^2. This is likely to be due to deflated estimates of error variance (0.54 and 0.97 for HapMap and CYP, respectively). The 2-SNP haplotypes were constructed by using each marker twice without taking into account the expected increases in the covariance between haplosets. This inflation of information may well lead to deflated error variances that exaggerate the significance level.

DISCUSSION

A popular belief is that haplotypes always provide greater power to detect disease genes when the SNPs tested are not functional but in LD with the causal locus. Interest in the analysis of haplotypes has increased as a result of the emphasis given by the International HapMap Project and other related initiatives [Salem et al., 2005]. Here we present evidence that haplotypes lead to poor estimates of localization and greater inconsistencies compared to single SNPs in an association analysis of 27 SNPs flanking the CYP2D6 gene. We have treated the simplest case, considering only haplosets of size 2 with haplotypes inferred for affected and normal from adjacent pairs of SNPs. The optimal number of SNPs in a haploset is not yet known. The simplest haploset of size 2 is particularly useful because intervals do not overlap (SNP intervals 1,2 and 2,3 do not overlap). However, even this simplest case of haplotypes has a drawback. With the exception of the first and the last markers in the region, all SNPs are used twice and this must account in part for the estimates of χ^2 being increased by approximately 60%. We attribute this outcome to the autocorrelation generated by the duplicated SNPs among haplosets, which deflates the error variance for association mapping. As a result, power is exaggerated and localization is poor compared to single SNPs. Longer windows of 3 or more SNPs overlap (e.g. 1,2,3 and 2,3,4), generating higher autocorrelations as the number of SNPs increases. Recent descriptions have focused on delimiting blocks of low haplotype diversity [Gabriel et al., 2002]. One option is to ignore overlapping windows and analyse haplotype blocks instead [Clark, 2004]. This approach will generate additional problems because block-finding algorithms are sensitive to marker density [Ke et al., 2004]. Block definition is arbitrary, and hence the beginning and end sites of a haplotype are unknown. To evade this problem, Lin et al. [2004] proposed sliding windows of all positions and lengths, but such designs will generate even higher autocorrelation because of the increased number of repeated SNPs. Furthermore, long haplotype windows have the problem of variable degrees of freedom, since different haplosets may yield different number of significant haplotypes. This design prohibits use of the z metric because of the difficulty in dichotomizing a large number of haplotypes by affection status, unless other metrics such as regression are used. If the choice is to use all haplotypes from the specific haploset, then the models will be over-parameterized and the tests computationally intensive. Longer haplotypes can be more ambiguous, especially in regions of recombination since they require phase information from diplotypes. Nevertheless, there are many ways that haplotypes could be used in association mapping, and powerful analytical tools have been proposed with favourable results [Morris, 2005]. In a comprehensive review of literature that is rapidly growing, Salem et al. [2005] showed that there are more than 40 published haplotyping methods. Our haplotype analysis did not provide better estimates of localization compared to single SNP tests, but the number of different ways to use haplotypes is limitless and alternative approaches that account for autocorrelation may obtain more favourable results.

Several authors have suggested that analysis of single SNPs loses power, especially for rare mutations [Lin et al., 2005], but that depends on the approach used for single SNP analysis. Association mapping is possible without an LD map, simply by selecting the most significant SNP. Single SNPs will not provide sufficient signal to narrow the region of interest, but this is also true when the most significant haplosets are selected. Hosking et al. [2002] observed significant associations with 14 SNPs, and the region of significance around CYP2D6 was reported to be 390 kb. Plotting the *P*-values on the kb map gives a pronounced "hole" in significance level for the CYP2D6 locus because two distant SNPs on either side of the gene are highly significant, making the surface bimodal (SNPs 12 and 18, Table I). Analysis of 2-SNP haplotypes also showed bimodality where the SNPs 17–18 haplotype ($\chi_1^2 = 58.5$) that contained the gene, was flanked by haplotypes on either side of the gene with greater χ_1^2 values (Table I). Hosking et al. [2002] and Meng

et al. [2003] also considered haplotypes in sliding windows of size 5. Although the levels of significance were considerably higher than single SNPs, haplotype analysis did not further refine the support interval. Selection of significant SNPs has the risk of losing information about other markers, and accepting a heavy correction that is unreasonable in a genome-wide scan. It further assumes that the functional SNPs have been included in the study. Although our multipoint approach is based on single SNPs, all association tests are considered simultaneously in a composite likelihood. Most importantly, it evades a heavy Bonferroni correction, giving a simple and powerful approach based on a 2×2 association table where the alleles of every SNP have been dichotomized by disease association (affected, normal). Single SNPs can be analysed in random samples, but also in cases and controls using an ascertainment correction based on the frequency of affected in the general population [Maniatis et al., 2005]. Data may be family-based where the affected offspring are the cases, while the non-transmitted alleles from the parents form pseudo-controls. The modelling procedures could employ various metrics other than z, (e.g. regression can be used for quantitative traits).

Some of the scepticism that surrounds the use of single SNPs arises because the methods do not consider that SNPs are in LD with one another. However, we have demonstrated that greater power and precision for localization are achieved with an LDU map, which describes the underlying LD structure. The SNPs in an LD map can be located in a block or in a step. Jeffreys et al. [2001] has shown that single-sperm genotyping can give recombination rate estimates at very fine scales, and Zhang et al. [2002] have shown that there is close agreement between LDU steps and sites of meiotic recombination. Sperm typing has two major limitations: it cannot determine female recombination, and can only be applied to small regions due to its high cost and effort. LDU maps on the other hand can be easily constructed for the entire genome using the data provided by the HapMap Project [Tapper et al., 2005]. These fine-scale metric maps provide valuable information about the pattern of LD. Using whole-chromosome linkage maps at low resolution, Tapper et al. [2005] have shown that more than 90% of the variation in LDU is explained by recombination. Low-resolution linkage maps are based on families with few meioses, whereas LDU maps reflect historical meiotic events. Although recombination dominates the LD structure, the great advantage of the LDU map method is that it models the decline of LD, which is due to pressures other than recombination, such as mutation, selection, and long-range migration.

The method presented here makes direct use of the HapMap data. The construction of HapMap and CYP maps yielded the same LDU length and similar structure. This paper examines for the first time the prospects for utilizing genome-wide LD maps for association studies in complex diseases. The results presented here provide evidence that small differences in the LDU map can influence the precision of association mapping. The fine-scale HapMap LDU map of chromosome 22 provided sufficient resolution to locate the CYP2D6 gene with a marginal error of 0.3 kb, compared to the 14.9 error that was obtained by mapping within the CYP LDU map. However, the HapMap LDU map did not reduce the confidence intervals from the original estimates based on the CYP map. The ultimate goal of the HapMap Project is more than three million SNPs [International HapMap Consortium, 2003]. We anticipate that the future HapMap releases and higher-resolution maps have the potential to further narrow the confidence interval.

Our approach for disease gene association mapping by LD is based on a model with evolutionary theory, which incorporates a parameter for the location of the causal polymorphism. When the locations of marker SNPs are expressed in LDU, then greater power was achieved to refine the location in the significant *CYP2D6* region. LDU maps ignore any phenotypic information. Their goal is to characterize the fine-scale pattern of LD in the genome. The main advantage of our proposed mapping approach is that it makes direct use of the HapMap data. Whether the objective is a genome-wide scan or a study on candidate regions, the HapMap LDU maps could be utilized directly and independently of the disease in question. Furthermore, the HapMap project has been developed for four different populations, Yoruba, Japanese, Chinese and CEPH, and therefore population-specific HapMap LDU maps offer versatility in association mapping studies. These population-specific LD maps have been incorporated in the Linkage Disequilibrium Location Data Base (LDDB), using genotype data for all four populations. LDDB also includes the most informative linkage map [Kong et al., 2004] together with the physical and cytogenetic maps.

ACKNOWLEDGMENTS

We are grateful to C.-F. Xu and L.K. Hosking from Discovery Genetics, GlaxoSmithKline, for the provision of the CYP2D6 SNP data and to the International HapMap Consortium for making the data publicly valuable. N.M. is grateful to Jane Gibson for her help with the HapMap LDU map.

WEB RESOURCES

The URLs for data presented herein are as follows:

The LDMAP program for the construction of LDU maps, http://cedar.genetics.soton.ac.uk/pub/PROGRAMS/LDMAP

LDDB, http://cedar.genetics.soton.ac.uk/public_html/

UCSC Genome Bioinformatics, http://genome.ucse.edu/.

REFERENCES

Abramowitz M, Stegun AI. 1964. Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables. New York: Dover Publications, Inc.

Beyer HW. 1966. Handbook of Tables for Probability and Statistics. Cleveland, OH: The Chemical Rubber Co.

Clark AG. 2004. The role of haplotypes in candidate gene studies. Genet Epidemiol 27:321–333.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Hill WG. 1974. Estimation of linkage disequilibrium in randomly mating populations. Heredity 33:229–239.

Hosking LK, Boyd RP, Xu CF, Nissum M, Cantone K, Purvis IJ, Khakhar R, Barnes MR, Liberwirth U, Hagen-Mann K, Ehm MG, Riley JH. 2002. Linkage disequilibrium mapping identifies a 390 kb region associated with CYP2D6 poor drug metabolising activity. Pharmacogenom J 2:165–175.

International HapMap Consortium. 2003. The International HapMap Project. Nature 426:789–796.

Jeffreys AJ, Kauppi L, Neumann R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29:217–222.

Ke X, Hunt S, Tapper WJ, Lawrence R, Stavrides G, Whittaker P, Collins A, Morris AP, Bentley D, Cardon LR, Deloukas P. 2004. Fine scale patterns of linkage disequilibrium across a 10 Mb region at 20q12-13.2. Hum Mol Genet 13:577–588.

Kong X, Murphy K, Raj T, He C, White PS, Matise TC. 2004. A combined linkage-physical map of the human genome. Am J Hum Genet 75:1143–1148.

Lin S, Chakravarti A, Cutler DJ. 2004. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. Nat Genet 36:1181–1188.

Lonjou C, Zhang W, Collins A, Tapper WJ, Elahi E, Maniatis N, Morton NE. 2003. Linkage disequilibrium in human populations. Proc Natl Acad Sci USA 100:6069–6074.

Maniatis N, Collins A, Xu C-F, McCarthy LC, Hewett DR, Tapper W, Ennis S, Ke X, Morton NE. 2002. The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc Natl Acad Sci USA 99: 2228–2233.

Maniatis N, Collins A, Gibson J, Zhang W, Tapper W, Morton NE. 2004. Positional cloning by linkage disequilibrium. Am J Hum Genet 75:846–855.

Maniatis N, Morton NE, Gibson J, Xu C-F, Hosking LK, Collins A. 2005. The optimal measure of linkage disequilibrium reduces error in association mapping *f* affection status. Hum Mol Genet 14:145–153.

Meng Z, Zaykin D, Xu C-F, Wagner M, Ehm MG. 2003. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. Am J Hum Genet 73: 115–130.

Morris AP. 2005. Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modelling of haplotypes. Genet Epidemiol 29:91–107.

Press HW, Teukolsky AS, Vetterling TW, Flannery PB. 1994. Numerical Recipes in C. The Art of Scientific Computing, 2nd edition. Cambridge: Cambridge University press.

Sachse C, Brockjmoller J, Bauer S, Roots I. 1997. Cytochrome 450 2D6 variants in a Caucasian population: allele frequencies and phenotype consequences. Am J Hum Genet 60:284–295.

Salem RM, Wessel J, Schork NJ. 2005. A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. Hum Genomics 2:39–66.

Tapper W, Collins A, Gibson J, Maniatis N, Ennis S, Morton NE. 2005. A map of the human genome in linkage disequilibrium units. Proc Natl Acad Sci USA 102:11835–11839.

Waldron ER, Whittaker JC, Balding DJ. 2006. Fine mapping of disease genes via haplotype clustering. Genet Epidemiol 30: 170–179.

Zhang W, Collins A, Maniatis N, Tapper W, Morton NE. 2002. Properties of linkage disequilibrium (LD) maps. Proc Natl Acad Sci USA 99:17004–17007.

APPENDIX 1

An $F(k,m)$ statistic follows an F distribution with k and m numerator and denominator degrees of freedom, respectively. In this study, k is the difference in numbers of parameters estimated between a null hypothesis and a more general hypothesis with m df, and m equals the number of SNPS minus the number of parameters estimated in the model. An $F(k,m)$ test is converted to a χ^2_1 in two steps: The first step obtains the corresponding probability P of $F(k,m)$ using an incomplete beta function based on a sub-routine given by Press et al. [1994, p 619]. The second step converts the χ^2_2 with 2 df, which is $-2\ln(P)$, to a χ^2_1 with 1 df using the Hastings approximation [Abramowitz and Stegun, 1965, p 933]. The table below shows results from our conversion procedures whereby tabulated F-values (upper critical values of the F distribution)

for various degrees of freedom and for two different levels of significance (95% and 99%) are converted to χ^2_1 estimates. These tabulated *F*-values can be found in most statistical textbooks. Here we give *F*-values from the tables presented by Beyer [1966], and show conversions from extreme values (e.g. $F_{1,1}$ or $F_{120,1}$) to accurate estimates of χ^2_2 and χ^2_1 (e.g. estimated values of χ^2_1 correspond precisely to tabulated χ^2_1 of 3.84 and 10.82 for 95% and 99%, respectively).

			Estimated		
Tabulated *F*-values	Numerator df (*p*)	Denominator df (n–p)	χ^2_2	χ^2_1	Significance level
161.40	1	1	5.991	3.843	95%
6.61	1	5	5.992	3.844	
4.96	1	10	5.988	3.840	
4.35	1	20	5.990	3.842	
3.92	1	120	5.991	3.843	
253.30	120	1	5.992	3.843	
405300.00	1	1	13.816	10.829	99%
47.18	1	5	13.815	10.829	
21.04	1	10	13.816	10.829	
14.82	1	20	13.816	10.830	
11.38	1	120	13.815	10.829	
634000.00	120	1	13.815	10.829	

For small *m* df, the conversion is more conservative compared to large *m*. For example, the tabulated value for $F_{1,120}$ approaches the estimated χ^2_1.