European Journal of Human Genetics (2015) 23, 1558–1565
© 2015 Macmillan Publishers Limited All rights reserved 1018-4813/15
www.nature.com/ejhg

ARTICLE

Kullback–Leibler divergence for detection of rare haplotype common disease association

Shili Lin*

Rare haplotypes may tag rare causal variants of common diseases; hence, detection of such rare haplotypes may also contribute to our understanding of complex disease etiology. Because rare haplotypes frequently result from common single-nucleotide polymorphisms (SNPs), focusing on rare haplotypes is much more economical compared with using rare single-nucleotide variants (SNVs) from sequencing, as SNPs are available and 'free' from already amassed genome-wide studies. Further, associated haplotypes may shed light on the underlying disease causal mechanism, a feat unmatched by SNV-based collapsing methods. In recent years, data mining approaches have been adapted to detect rare haplotype association. However, as they rely on an assumed underlying disease model and require the specification of a null haplotype, results can be erroneous if such assumptions are violated. In this paper, we present a haplotype association method based on Kullback–Leibler divergence (hapKL) for case–control samples. The idea is to compare haplotype frequencies for the cases *versus* the controls by computing symmetrical divergence measures. An important property of such measures is that both the frequencies and logarithms of the frequencies contribute in parallel, thus balancing the contributions from rare and common, and accommodating both deleterious and protective, haplotypes. A simulation study under various scenarios shows that hapKL has well-controlled type I error rates and good power compared with existing data mining methods. Application of hapKL to age-related macular degeneration (AMD) shows a strong association of the complement factor H (*CFH*) gene with AMD, identifying several individual rare haplotypes with strong signals.

European Journal of Human Genetics (2015) **23**, 1558–1565; doi:10.1038/ejhg.2015.25; published online 4 March 2015

INTRODUCTION

Rare variants have been investigated intensely in the past few years for their association with common diseases.[1] The type of rare variants that has been the main focus is single-nucleotide variants (SNVs), owing to the development and rapid availability of next-generation sequencing (NGS) technology. However, there is a continuing realization that rare haplotype variants resulting from common SNVs may also have an important role in understanding complex disease etiology.[2–10] The interest in detecting rare haplotype association with common diseases is further fueled by the recognition that rare haplotype may tag rare causal SNVs.[7–10] There are advantages pursuing rare haplotypes instead of rare SNVs. First, rare haplotypes frequently result from common single-nucleotide polymorphisms (SNPs); therefore, focusing on detection of rare haplotype association is much more economical compared with using rare SNVs from NGS, as SNPs are available and 'free' from already amassed genome-wide association studies (GWAS). Further, associated haplotypes may shed light on the underlying disease causal mechanism, a feat unmatched by SNV-based collapsing methods.

A number of new statistical methods have been proposed specifically to detect rare haplotype association because existing methods in GWAS are typically not amendable to rare haplotypes.[11] For example, generalized linear model (GLM)-based methods[12–14] may encounter non-convergence in its expectation-maximization (EM) estimates when challenged with rare haplotypes. Among such new approaches, the majority use likelihood-based regularization methods (eg, Lasso[15]) to weed out unassociated haplotypes[3–5,7,8] so that those that are associated with the disease, especially the rare ones, can be more precisely estimated for their effects on the trait. However, owing to the difficulty in evaluating the effect of the uncertainty of regularization parameters on assessing association, the Bayesian counterpart of Lasso has been proposed for studying rare haplotype association,[6,9,10] as well as the Bayesian hierarchical GLM approach.[5]

Regardless of whether a method is likelihood based or Bayesian formulated, such a method relies on assuming an underlying model connecting the haplotypes to the disease, which, unfortunately, is unknown. Theoretically, it is possible to consider multiple hypothesized models, and then either choose the most likely one or perform model averaging, but such an approach will increase their computational intensity. Although tests for overall haplotype association using common SNPs have been shown to be more powerful than rare SNV collapsing methods,[7,8] a major advantage of regularized haplotype association methods is its ability of detecting individual haplotype associations to shed light on underlying disease causal mechanisms. However, for individual haplotype detection methods, it is necessary to specify a base haplotype: a null haplotype that is not associated with the disease. This is a major drawback as such a haplotype can be elusive. Although the reference haplotype is typically assumed to be the one with the largest frequency, such an assumption can be erroneous and can lead to incorrect interpretation of results.[6]

Departing from the model-based methods discussed above is a two-step, model-free, approach,[16] which is simple and easy to implement. However, the method is susceptible to the existence of both 'risk' and 'protective' haplotypes; the approach will lose power as the effects of

Department of Statistics, The Ohio State University, Columbus, OH, USA
*Correspondence: Professor S Lin, Department of Statistics, The Ohio State University, 1958 Neil Avenue, 404 Cokins Hall, Columbus 43210-1247, OH, USA. Tel: +1 614 292 7404; Fax: +1 614 292 2096; E-mail: shili@stat.osu.edu

Received 23 June 2014; revised 16 November 2014; accepted 28 January 2015; published online 4 March 2015

these two types of haplotypes will cancel out one another.[6,11] In this paper, we attempt to reap the benefits of a simple nonparametric approach without the need to specify a null haplotype, yet being able to accommodate both risk and protective variants. Specifically, we propose a model-free approach for detecting haplotype association based on Kullback–Leibler divergence (hapKL) for case–control samples. The idea is to compare haplotype frequencies for the cases and the controls by computing symmetrical divergence measures. In addition to being able to deal with haplotypes having effects of opposite directions, another important property of such measures is that both the frequencies and the logarithms of the frequencies contribute in parallel, thus balancing the contributions from the rare and the common haplotypes. Further, this framework allows us to carry out both an overall test as well as tests for individual haplotype effects. Most importantly, unlike model-based approaches, specification of a null haplotype as a base haplotype is unnecessary. To validate hapKL, we evaluate and compare it with a couple of existing methods by carrying out a simulation study under various scenarios. We also apply hapKL to a data set studying the genetic basis of age-related macular degeneration (AMD). The result shows a strong association of the complement factor H (*CFH*) gene with AMD, replicating previously reported results. More importantly, hapKL also reveals several individual haplotype associations, including not only a rare one found in previous studies but also a couple other rare haplotypes with strong signals. Most importantly, this real data analysis, in which a ‘null’ haplotype is not known *a priori* (unlike simulation studies), provides a good example to demonstrate the advantage of hapKL compared with model-based approaches.

Despite the fact that sequencing data are gradually becoming available in public databases such as dbGap (http://dbgap.ncbi.nlm.nih.gov), such data are still much more limited compared with the massive amount of already existing GWAS common SNP data, which constitute a treasure trove, but unfortunately are still underutilized and underanalyzed. The use of such data not only is economical (as no additional genotyping is needed) but also is scientifically advantageous in some cases as rare haplotypes constructed based on GWAS common SNPs may tag multiple rare causal SNVs without resequencing.[7,8] Further, it has already been shown in the literature that haplotype methods can be more powerful for detecting such rare variant association compared with collapsing methods[7,10] (hence, no direct comparison of hapKL with collapsing methods are carried out in this work). Most importantly, haplotype methods can identify specific associated haplotypes and SNPs, thus making them informative for designing further experiments and studies to understand causal mechanisms. In this work, we will use the AMD common SNP data to illustrate the power of using common SNPs to detect rare variant association. This is an especially good benchmark data set for evaluating nonparametric methods such as hapKL, as rare associated haplotypes have been successfully detected by modeling-based methods.[6,9]

MATERIALS AND METHODS

The hapKL measures

In probability or information theory, the KL divergence,[17] more popularly known as relative entropy in computer science, is a nonsymmetric measure of the ‘distance’ between two probability distributions F and G. Specifically, for discrete distributions, the KL divergence of $G=\{G_i,\ i=1,\ \ldots\}$ from $F=\{F_i,\ i=1,\ \ldots\}$ is defined as: $\sum_i F_i \log(F_i/G_i)$, which basically measures the expected value of the logarithmic difference between F and G based on the (true) probability distribution F. In our setting, neither the frequency distribution among the cases nor that among the controls is the ‘true’ probability distribution. Therefore, we consider a symmetrized measure as defined in the following.

Suppose there are a total of k haplotypes present in the case–control sample. Let $F=\{F_1, \ldots, F_k\}$ and $G=\{G_1, \ldots, G_k\}$ be the haplotype frequency distributions among the cases and controls, respectively, where F_i and G_i are the frequencies of haplotype h_i. The KL divergence for an overall association measures the symmetrical difference:

$$\text{hapKL}_{all} = \frac{1}{2}\left(\sum_{i=1}^{k} F_{\mathrm{i}} \log\frac{F_{\mathrm{i}}}{G_{\mathrm{i}}} + \sum_{i=1}^{k} G_{\mathrm{i}} \log\frac{G_{\mathrm{i}}}{F_{\mathrm{i}}}\right) \equiv \frac{1}{2}\left(\sum_{i=1}^{k} (F_{\mathrm{i}} - G_{\mathrm{i}}) \log\frac{F_{\mathrm{i}}}{G_{\mathrm{i}}}\right)$$

It is important to note that, in the above measure, both the haplotype frequencies and the logarithms of the frequencies contribute in parallel, thus balancing the contributions from the rare and the common haplotypes, making it appropriate for detecting rare haplotype–disease association. The second expression also facilitates the observation that effects of both directions (deleterious and protective) can be accommodated. To see this more clearly, suppose h_i is a rare haplotype with $F_i > G_i$, then $F_i - G_i$ will be small (positive), whereas $\log(F_i/G_i)$ will typically be large (also positive). On the other hand, suppose h_j is a common haplotype with $F_j < G_j$, then $F_j - G_j$ will be large (but negative), whereas $\log(F_j/G_j)$ will typically be small (also negative). This will make both $(F_i - G_i)\ \log(F_i/G_i)$ and $(F_j - G_j)\ \log(F_j/G_j)$ contribute fairly stable positive values to hapKL_{all}. Therefore, rare and common haplotypes, as well as haplotypes with opposite effects, will all contribute to the test statistic synergistically. Another desirable property of hapKL is that there is no need to specify a reference, or base, haplotype. This is a major advantage since setting the right base haplotype is a tricky issue.[6] However, we note that upper bound does not exist for hapKL;[18] this point will be elaborated in the Discussion section.

In addition to the overall measure of association, it is of importance to investigate the effects of individual haplotypes on the disease. To achieve this goal, we consider the individualized measure of association for haplotype h_i:

$$\text{hapKL}_i = \frac{1}{2}\left(F_{\mathrm{i}} \log\frac{F_{\mathrm{i}}}{G_{\mathrm{i}}} + G_i \log\frac{G_{\mathrm{i}}}{F_{\mathrm{i}}}\right), \quad i = 1, \cdots, k,$$

which basically measures the discrepancy of the frequency of haplotype h_i between the cases and the controls. Thus, the KL divergence framework allows us to devise not only an overall statistic but also individualized statistics to test for association of specific haplotypes. The latter is extremely useful for detecting epistatic interactions of multiple variants. Such results can shed light on the underlying causal mechanisms and can aid in designing experiments to validate findings.

Assessment of significance

It can be easily seen that hapKL_{all} and hapKL_i are non-negative measures. When $F=G$, $\text{hapKL}_{all}=0$ and $\text{hapKL}_i=0$ for any haplotype h_i. Further, the more discrepant the two distributions are, the larger the measure. Therefore, a divergence significantly larger than expected under the null hypothesis means significant association. Specifically, we use a Monte Carlo simulation technique to assess whether there is a significantly large divergence. Treating all observed data as a whole set, we randomly divide the set as case and control samples to match the numbers in the original data. We then compute $\text{hapKL}_{all}^{b}\ \ b = 1, \ldots,\ B$, based on a total of B random samples to build the null distribution. The P-value is computed as $\sum_{b=1}^{B} I\left(\text{hapKL}_{all} < \text{hapKL}_{all}^{b}\right)/B$, where $I(\cdot)$ is the usual indicator function taking the value of 1 or 0 depending on whether the condition within the parentheses is satisfied or not. The P-value for an individual haplotype can be computed similarly.

Computing haplotype frequency distributions

Our data are from GWAS SNPs, and therefore we only have genotypes, not haplotype, data. Although haplotype may be inferred from genotypes, this is the exception rather than the rule. Fortunately, hapKL only requires the frequency distributions, not the individual haplotypes. EM algorithm is the most effective, computationally efficient and frequently used method to estimate haplotype frequency distributions based on SNP data. In particular, we use the R package haplo.stats[12] to compute the haplotype frequencies without filtering out rare haplotypes. We perform this estimation separately for the case and control samples to obtain the frequency distributions F^* (for the cases) and G^* (for the controls). To compute the hapKL measures, F and G are required to have the same support. To achieve this, we let F and G be defined on the union of the haplotypes appearing in F^* and G^*. For any haplotype that only appears in G^*, we allow for a very small frequency (say, $\leq 10^{-6}$) in F, and *vice versa*. Finally, F and G are renormalized to be proper probability distributions before they are used in the hapKL calculation.

Software

The methods presented in this paper have been implemented in R, which can be downloaded freely from http://www.stat.osu.edu/~statgen/SOFTWARE/hapKL.

RESULTS

Simulation study

We carried out a simulation to thoroughly evaluate the performance of hapKL, and to compare it with LBL[6] and hapassoc,[14] both logistic regression modeling based. We perform three sets of simulation: the first is to gauge the type I error rate of hapKL; the second is to compare the performance with LBL and hapassoc; the last simulation is to show that hapKL performs well when the underlying disease model is not additive, a departure from the typical assumption in LBL and hapassoc. To make it easier to compare with LBL and hapassoc, we consider the same three haplotype settings that were used to compare LBL with hapassoc,[6] but with additional disease settings so that not only type I error for individual haplotypes, but overall type I error rates, can be ascertained to study hapKL more thoroughly.

Table 1 gives three haplotype settings (HS1, HS2 and HS3), that is, haplotype distributions, with 6, 9 and 12 haplotypes, respectively. In each setting, there are two rare haplotypes (frequency ≤ 0.01) for devising various models to study the performance of hapKL for handling rare haplotype associations. Specifically, our first set of simulation considers a null model in which all the odds ratios (against the unassociated reference haplotype, the last haplotype in each setting) are set to be 1. The second set of simulation considers three additive disease models: RR (rare-rare – both rare haplotypes are causal with odds ratios of 3 and 2), RC (rare-common – one of the rare haplotypes and one common haplotype are causal, also with odds ratios of 3 and 2) and C (common – only one common haplotype is causal with an odds ratio of 2). Our third, and last, set of simulation entertains a model in which both rare haplotypes are causal (the RR model), but the causal effect of the less rare haplotype exerts a dominant effect on the disease (ie, one or two copies of the haplotype will have the same effect on the disease), leading to the nonadditive RR model. The purpose of this simulation is to evaluate the robustness of hapKL to departure from additive model, and thus we do not present results for LBL or hapassoc. For each haplotype setting and disease model combination given in Table 1, we consider three sample sizes, 400, 1000 and 2000 individuals, with an equal number of cases and controls. We set the number of random simulations B to be 500 in all computation of P-values.

Type I error. We first present the simulation results under the null model to gauge the type I error rate of hapKL, both in terms of the overall association test and as tests for the individual haplotypes. Remember that for each sample simulated from the null setting, 500 permuted samples are obtained by permuting the affection status to build the null distribution for computing the P-value, as described in Materials and methods section. As such, the empirical type I error rate is not guaranteed to be exactly the same as the nominal type I error rate. Table 2 shows the percentage that a test is rejected (for overall association and for each of the haplotypes). Provided in the table is also a column named ‘Ave’, which provides the average of the type I errors over all individual haplotypes. As one can see from the table, the type I errors for the overall association and the average over individual haplotypes under various haplotype settings and sample sizes are all below 5%. The type I error rates for individual haplotypes vary a bit around 5% (both above and below), which is expected given the moderate number of simulations for computing the P-values and the fact that there are haplotypes with rather small frequencies.

Table 1 Haplotype settings (frequency distributions) and odds ratios under various association scenarios[a]

Haplotype setting	Hap	Freq	Null model	*Additive models*			Nonadditive RR
				RR	RC	C	
HS1	01100	0.3	1	1	1	1	1
	10100	0.005	1	3	3	1	3
	11011	0.01	1	2	1	1	2[b]
	11100	0.155	1	1	1	1	1
	11111	0.11	1	1	2	2	1
	10011	0.42	1	1	1	1	1
HS2	01010	0.06	1	1	1	1	1
	01100	0.25	1	1	1	1	1
	10000	0.08	1	1	2	2	1
	10100	0.005	1	3	3	1	3
	11011	0.01	1	2	1	1	2[b]
	11100	0.09	1	1	1	1	1
	11101	0.085	1	1	1	1	1
	11111	0.1	1	1	1	1	1
	10011	0.32	1	1	1	1	1
HS3	00111	0.07	1	1	1	1	1
	01000	0.02	1	1	1	1	1
	01011	0.05	1	1	1	1	1
	01101	0.06	1	1	1	1	1
	01110	0.14	1	1	1	1	1
	10010	0.08	1	1	2	2	1
	10100	0.005	1	3	3	1	3
	11011	0.01	1	2	1	1	2[b]
	11101	0.09	1	1	1	1	1
	11110	0.13	1	1	1	1	1
	11111	0.1	1	1	1	1	1
	10001	0.245	1	1	1	1	1

[a] Under the scenario where all odds ratios are 1 (null model), none of the haplotypes is associated with the disease, and this scenario is used to assess the overall type I error of hapKL. The three association scenarios are RR (rare-rare – two rare associated haplotypes), RC (rare-common – one rare and one common associated haplotypes) and C (common – one common associated haplotype). They are mainly used to evaluate power; however, they can also be used to assess the type I errors for individual haplotypes not associated with the disease.
[b] The effect of this haplotype under the nonadditive RR model acts dominantly. That is, an individual carrying two copies of this haplotype has the same effect on the disease as an individual carrying one copy and another haplotype that does not have an effect on the disease.

Power comparison. Our second set of simulation is to compare the performance of hapKL with LBL and hapassoc in terms of power. We also, as an aside, compare type I error rates on individual haplotypes that are not associated with the disease. Recall that hapKL does not rely on the assumption of an underlying model, whereas both LBL and hapassoc are model based. Therefore, this study is mainly to investigate whether hapKL is underpowered given its nonparametric

Table 2 Empirical type I error rate (percentage) at 5% nominal level for the null model

				Individual haplotypes[a]											
Hap setting	*Sample size*	*Overall assoc* [b]	*Ave* [c]	*h1*	*h2*	*h3*	*h4*	*h5*	*h6*	*h7*	*h8*	*h9*	*h10*	*h11*	*h12*
HS1	400	3	2.67	3	2	3	2	2	4						
	1000	3	4.5	7	2	2	5	6	5						
	2000	2	4	4	4	3	4	4	5						
HS2	400	4	4	5	1	5	4	4	7	7	2	1			
	1000	4	4.11	3	7	7	1	4	4	6	1	4			
	2000	2	3.78	2	2	5	5	4	5	3	5	3			
HS3	400	1	3.08	4	2	4	3	5	3	6	1	4	2	1	2
	1000	3	3.92	5	7	4	4	2	4	3	5	3	5	3	2
	2000	3	3.6	2	6	2	3	0	7	1	4	4	5	6	4

[a] Notation for haplotypes for each setting, with those having frequencies <0.05 (rare haplotypes) within parentheses: HS1 – h1=h01100, h2=(h10100), h3=(h11011), h4=11100, h5=11111, h6=10011; HS2 – h1=01010, h2=01100, h3=10000, h4=(10100), h5=(11011), h6=11100, h7=11101, h8=11111, h9=10011; HS3 – h1=00111, h2=01000, h3=01011, h4=01101, h5=01100, h6=10010, h7=(10100), h8=(11011), h9=11101, h10=11110, h11=11111, h12=10001. [b] Type I error rate for overall association for the region spanned by the five SNPs. [c] The average type I error rate averaging over the type I error rates for all individual haplotypes.

nature. For this simulation study, the assumed disease model (additive model, as it is typically the default) is the same as the simulation model, thus giving an advantage to the two model-based approaches. Further, both LBL and hapassoc require a reference haplotype, which is also set to be the true reference haplotype in the simulation. The results for the data simulated under the additive RR model are given in Figure 1. As we can see from the figure, for all combination of haplotype settings and sample sizes, all three methods have well-controlled type I error rates for the haplotypes not associated with the disease, fluctuating around the 5% nominal significance level marked by the solid horizontal line. It can be seen that hapKL has greater power than LBL and hapassoc for detecting the association of both rare haplotypes for almost all haplotype settings and sample sizes, except for the rarer haplotype under setting 3 when the sample size is small. The results for the data simulated under the additive RC model are given in Figure 2. For detecting the association with the common haplotype, all three methods performed similarly, although hapKL is slightly better for several of the setting and sample size combinations. However, hapKL is being out-performed for detecting the rare haplotype association under haplotype setting 1 when the sample sizes are large (1000 and 2000). Finally, the results for data simulated under the additive C model are given in Figure 3. All three methods perform almost identically for detecting the only associated common haplotype, but hapKL is clearly more powerful when the sample size is small (400). All the results for LBL and hapassoc presented in the figures are the same as in Biswas and Lin.[6] Note that the type I error for LBL is determined based on the Bayes factor, whereas the type I error for hapassoc was obtained by setting an appropriate *P*-value cutoff so that its type I error is similar to that of LBL for comparison purpose. Further details can be found in Biswas and Lin.[6] In summary, even though LBL and hapassoc assume the true underlying model for haplotype effects and true reference haplotype, hapKL, being a nonparametric approach, outperforms them in some of the scenarios and has comparable power/type I error in the other scenarios investigated, without either assumptions.

Nonadditive RR model. To demonstrate that hapKL still performs well when the underlying model is nonadditive, we considered the RR scenario in which both associated haplotypes are rare but the effect of the haplotype with an odds ratio of 2 does not confer risk additively. The results, given in Table 3, show that the average type I errors across all nonassociated haplotypes is well under control at around the nominal level of 5% for all sample sizes (rows with odds ratio being 1). When the sample size is 2000, there is considerable power for detecting the haplotype with an odds ratio of 3, even though it is rarer compared with the other associated haplotype, similar to earlier results (Figure 1) when additive models were considered.

Application to AMD

AMD, which affects tens of millions of elderly individuals worldwide, is a late-onset common eye condition that can lead to vision loss and even blindness. GWAS have identified several SNPs in the *CFH* gene on chromosome 1 to be associated with AMD.[19] Haplotype analyses have also been carried out using GWAS SNPs spanning the *CFH* gene in an attempt to discern the causal mechanism.[6,11,20] In particular, Spencer *et al*[11] identified three 8-SNP common haplotypes, one confers strong risk effect on, and two are protective against, AMD using a traditional score test.[12] They also reported deleterious effect of a pooled variant (pooling over all rare haplotypes resulted from the eight SNPs). On the other hand, Biswas and Lin[6] used LBL to perform haplotype analysis so that the effects of individual rare haplotypes can be ascertained. They considered a region spanned by seven SNPs implicated in previous studies, five from Spencer *et al*[11] and two from Klein *et al.*[19] These seven SNPs are, in their chromosomal order, rs3753394 (chr1.hg19:g.196620917C>T), rs800292 (chr1.hg19:g.196673103G>A), rs203674 (chr1.hg19:g.196684625G>T), rs3753396 (chr1.hg19:g.196695742A>G), rs380390 (chr1.hg19:g.196701051G>C), rs1329428 (chr1.hg19:g.196702810C>T) and rs1065489 (chr1.hg19:g.196709774G>T), with the fifth and the sixth SNPs reported in the Klein study[19] and the remaining from the Spencer study.[11] The sample analyzed by LBL consisted of 315 cases and 149 controls. The SNPs were genotyped on the Affymetrix 100K and Illumnina 100K platforms, and were made available by the National Eye Institute through dbGaP (http://www.ncbi.nlm.nih.gov/gap; phs000001.v3.p1). Taking haplotype TGTGCCT as the base (neutral) haplotype (same haplotype used in Spencer *et al*[11]), LBL identified three risk haplotypes: the most common one (CGGAGCG) matches up with the nucleotides of the five SNPs (CGGA–G) on the risk haplotype identified by Spencer *et al*;[11] the other two are TGGAGCG (close to being rare; overall frequency=0.075) and CGTGCCT (rare haplotype;

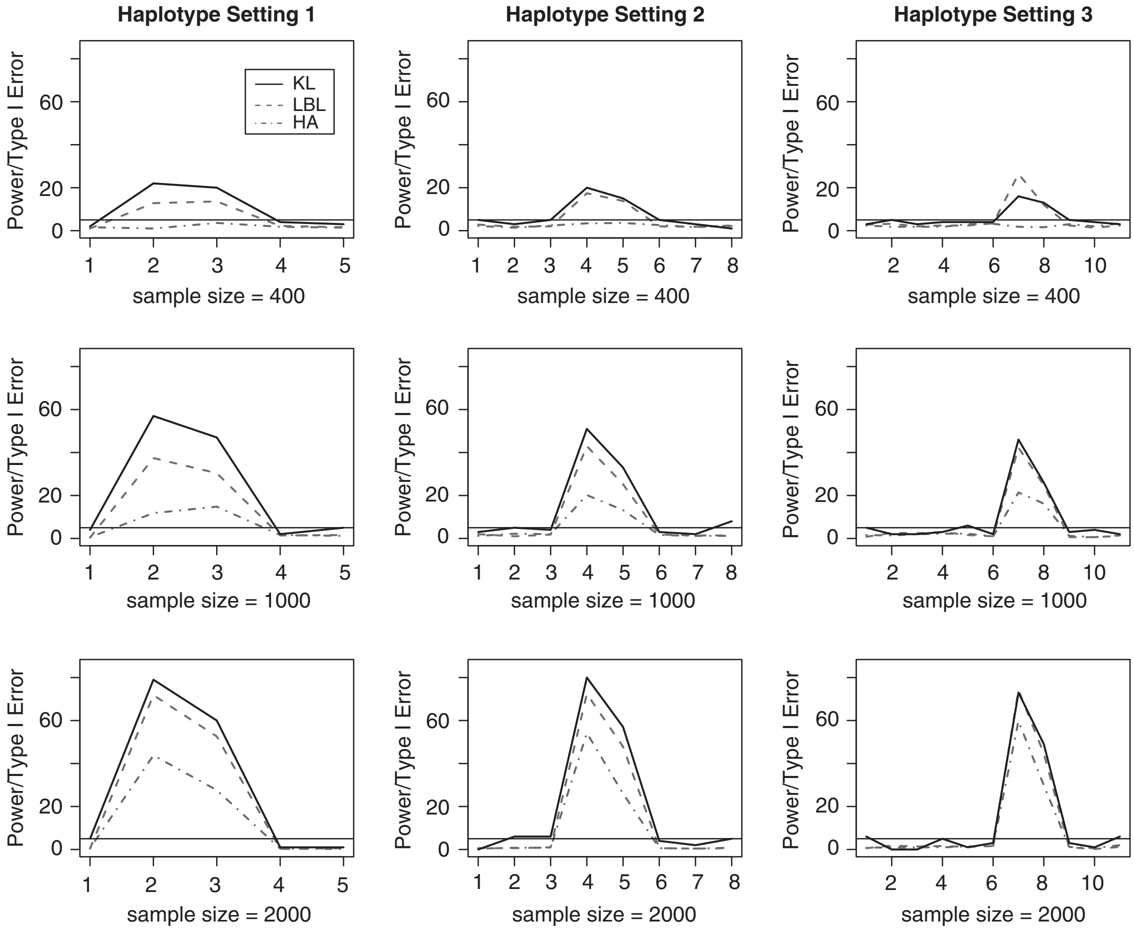

Figure 1 Power/type I error (percentage) for detecting individual haplotype associations under the RR (two rare associated haplotypes) model. The arrangement of haplotypes on the x axis is provided below. Note that the specification of a reference haplotype is necessary for LBL and hapassoc (HA) (but not for hapKL). As a consequence, even though there are 6, 9 and 12 haplotypes in haplotype setting 1 (HS1), haplotype setting 2 (HS2) and haplotype setting 3 (HS3), respectively, power/type I error are plotted for only 5, 8 and 11 haplotypes. This is also true for Figures 2 and 3. HS1: (1, 2, 3, 4, 5)=(01100, 10100, 11011, 11100, 11111), with haplotype 10011 treated as the reference haplotype and hence not appearing in the figure. Further, haplotypes 1, 4 and 5 are not associated the disease. HS2: (1, 2, 3, 4, 5, 6, 7, 8)=(01010, 01100, 10000, 10100, 11011, 11100, 11101, 11111), with haplotype 10011 treated as the reference. Further, haplotypes 1–3 and 6–8 are not associated with the disease. HS3: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)=(00111, 01000, 01011, 01101, 01110, 10010, 10100, 11011, 11101, 11110, 11111), with 10001 as the reference. Further, haplotypes 1–6 and 9–11 are not associated with the disease.

overall frequency=0.011). They further identified a rare protective haplotype TGTACTG (overall frequency=0.046), whose nucleotides at the fifth and sixth SNPs (——CT-) are complementary to the two risk variants identified by Klein *et al.*[19] Further details on the haplotypes can be found in the literature.[11,19,20]

We applied hapKL, with 10000 random simulations, to the same data set analyzed by LBL, but without setting any base haplotype to be more objective, and an interesting picture emerged. First, the so-called 'neutral' haplotype TGTGCCT assumed in the LBL analysis turns out to be marginally significant (*P*-value=0.065) as a protective haplotype (Table 4). (Note that this haplotype was selected as 'neutral' haplotype in the LBL analysis[6] because it is not rare and the frequencies for the cases and controls are similar.) As such, the effects of other haplotypes that are also protective would be reduced in the LBL analysis and might be missed. Indeed, in addition to haplotype TGTACTG, hapKL also identified three other protective haplotypes (CGTACTG, CATACTG and TGTGCCG), with the last one being rare (Table 4). The nucleotides of the five SNPs from the Spencer study[11] in the two common protective haplotypes, in fact, are the same as the corresponding nucleotides in the two common associated haplotypes reported there. For risk haplotypes, hapKL also identified CGGAGCG to be highly significant, consistent with the results from the earlier studies.[6,11] The rare haplotype, CGTGCCT, identified by LBL, is also implicated by hapKL to be marginally significant. On the other hand, haplotype TGGAGCG has a *P*-value of 0.0959 from hapKL, although LBL's results indicated that it was highly significant. These results are not surprising because the effect of a risk-leaning haplotype would typically be magnified using LBL if a protective (leaning) haplotype was assumed to be neutral. An additional risk haplotype identified by hapKL is TGTACTT, which is extremely rare with haplotype frequency in the cases and controls estimated to be 0.01 and 0, respectively. In addition to individual haplotypes, hapKL also provided the result for an overall test of association, which is extremely significant with a *P*-value of 0.0001.

DISCUSSION

Detecting uncommon causal variants with common SNPs has gained increasing attention because it has the potential to detect rare causal

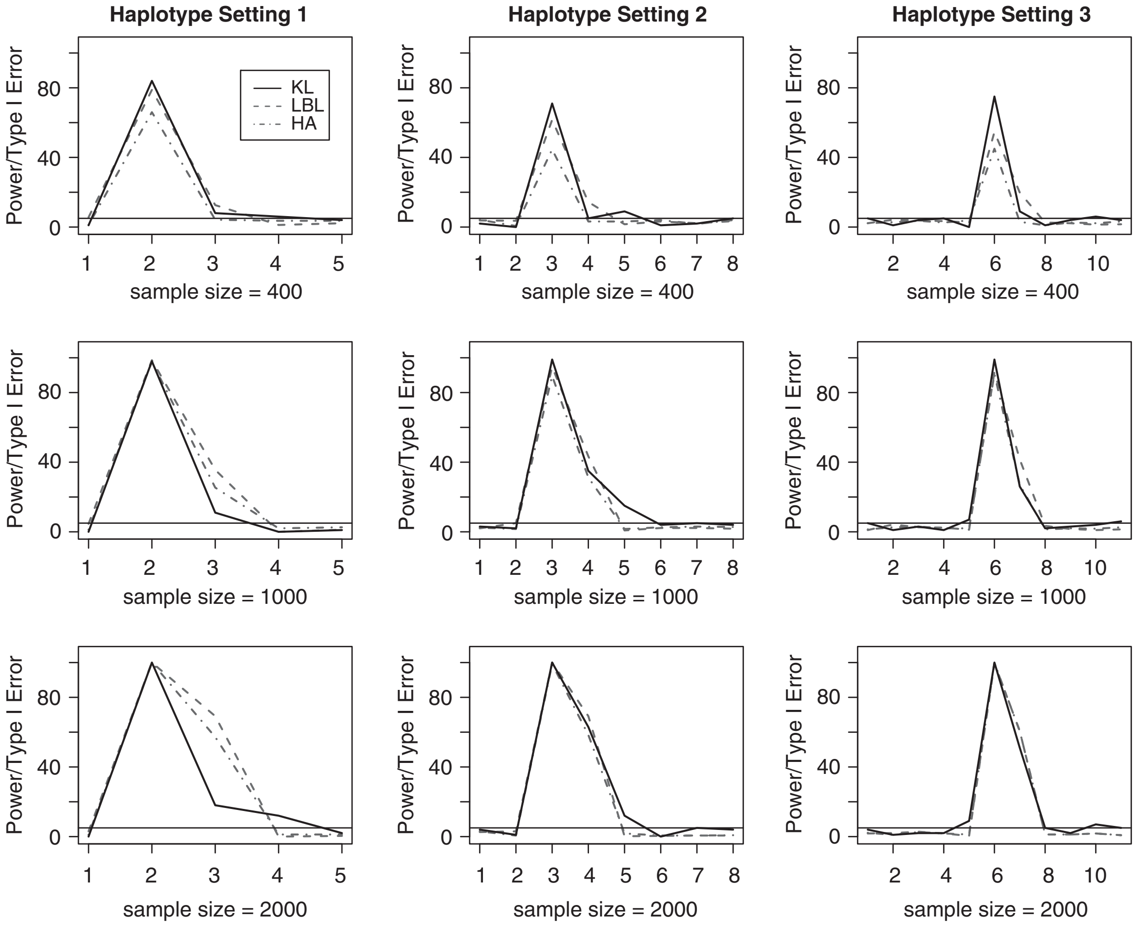

Figure 2 Power/type I error (percentage) for detecting individual haplotype associations under the RC (one rare and one common associated haplotypes) model. The arrangement of haplotypes on the x axis is as follows: HS1 – (1, 2, 3, 4, 5) = (11011, 11111, 10100, 01100, 11011), with haplotype 10011 treated as the reference haplotype and hence not appearing in the figure. Further, haplotypes 1, 4 and 5 are not, whereas halotypes 2 (common) and 3 (rare) are, associated with the disease. HS2 – (1, 2, 3, 4, 5, 6, 7, 8) = (01010, 11011, 10000, 10100, 01100, 11100, 11101, 11111), with haplotype 10011 treated as the reference. Further, haplotypes 1, 2 and 5–8 are not, whereas haplotypes 3 (common) and 4 (rare) are, associated with the disease. HS3 – (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11) = (00111, 01000, 01011, 01101, 01110, 10010, 10100, 11011, 11101, 11110, 11111), with 10001 as the reference. Further, haplotypes 1–5 and 8–11 are not, whereas halotypes 6 (common) and 7 (rare) are, associated with the disease.

variants without the need to perform sequencing-based regenotyping. Departing from the majority of existing methods that are model based, we propose hapKL in this paper, which is model free and is based on measuring the distance between the cases and the controls in terms of the distributions of haplotypes or individual haplotype using relative entropy. This approach can detect both overall association and associations of individual haplotypes with deleterious or protective effects in a single analysis instead of only focusing on one of them as in existing methods, thereby increasing its versatility. Our simulation study shows that hapKL has well-controlled type I error and its model-free nature does not render it underpowered compared with two model-based methods, LBL and hapassoc, even when the analysis model was set to be the same as the simulation model for LBL and hapassoc. Furthermore, because hapKL is nonparametric, the results are not sensitive to the underlying disease model, and there is no need to assess the effect due to model uncertainty. However, given the simulation is limited in scope, we by no means claim superiority of hapKL to model-based approaches in general. In the contrary, hapKL is unable to accommodate covariates, which may lead to inflated type I error if confounders exist, as we discussed below. It would be of interest to compare all rare variant haplotype association approaches, both model based and model free, to reach a more definitive conclusion. However, this is out of the scope of the current paper, as our main goal of this contribution is to demonstrate the advantage of hapKL without the need to specifying a base haplotype, and to draw attention to the potential problem from a model-based approach had a baseline haplotype been specified incorrectly, as we discuss next.

HapKL is free of the concept of a reference haplotype, the most important advantage compared with its model-based counterparts. Selection of a reference haplotype is an extremely tricky issue; as we saw in our analysis of the AMD data, the selection of a reference haplotype is a subjective decision and can have a great impact on the outcomes. The default reference haplotype, as in the LBL software, is typically set to be the one with the largest frequency, but this can have a detrimental effect if such a haplotype is either a deleterious or a protective one. This was correctly recognized in the earlier studies;[6,11] as the most frequent haplotype (CGGAGCG) was a risk one for AMD, the base haplotype in the LBL analysis was set to be a different haplotype (TGGSGCG). Nevertheless, even if a haplotype is carefully selected to minimize its chance of being an associated haplotype, its neutrality is not guaranteed as we saw in the AMD example. In general, if a protective-leaning haplotype is selected as the reference

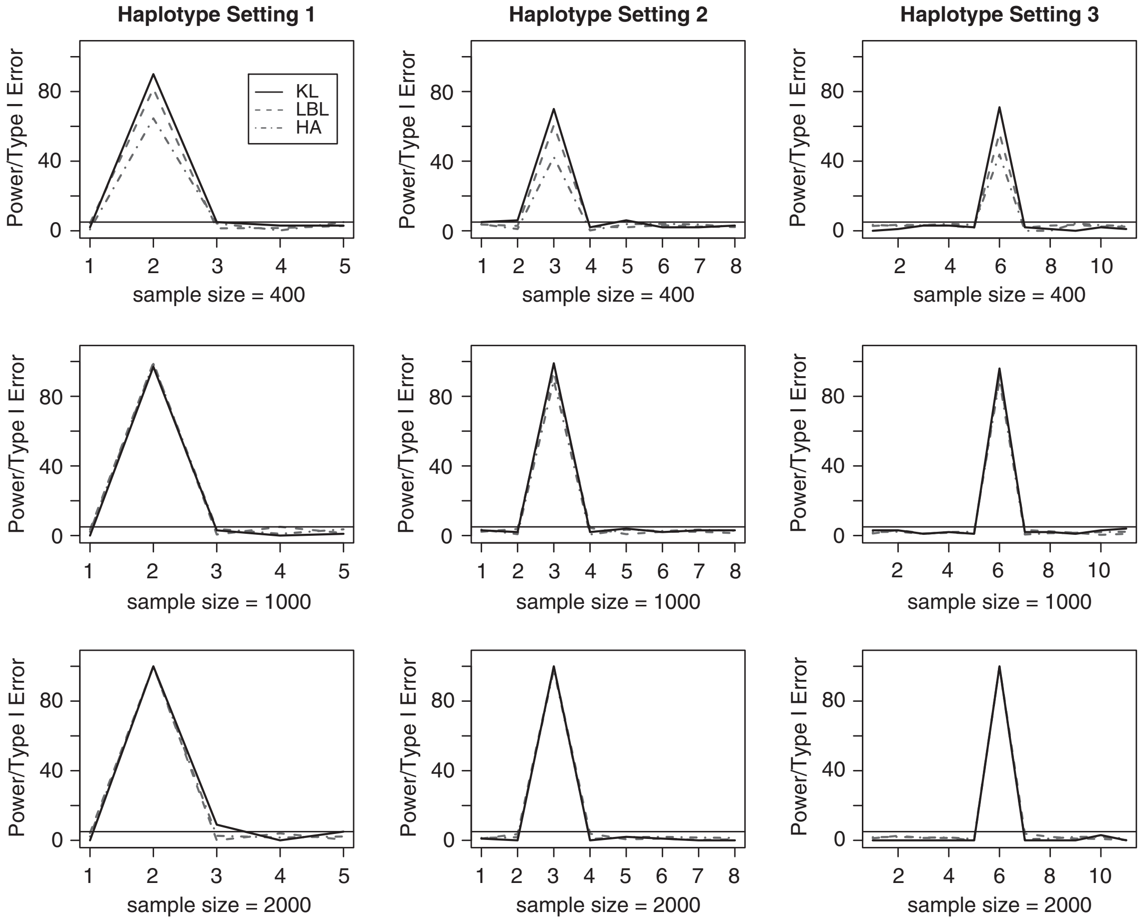

Figure 3 Power/type I error (percentage) for detecting individual haplotype associations under the C (one common associated haplotype) model. The arrangement of haplotypes on the x axis is as follows: HS1 – (1, 2, 3, 4, 5) = (11011, 11111, 01100, 10100, 11011), with haplotype 10011 treated as the reference haplotype and hence not appearing in the figure. Further, haplotypes 1 and 3–5 are not associated with the disease. HS2 – (1, 2, 3, 4, 5, 6, 7, 8) = (01010, 11011, 10000, 10100, 01100, 11100, 11101, 11111), with haplotype 10011 treated as the reference. Further, haplotypes 1, 2 and 4–8 are not associated with the disease. HS3 – (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11) = (00111, 01000, 01011, 01101, 01110, 10010, 10100, 11011, 11101, 11110, 11111), with 10001 as the reference. Further, haplotypes 1–5 and 7–11 are not associated with the disease.

Table 3 Power[a] and type I error[b] (percentage) at 5% nominal level for a model with nonadditive haplotype effects

			Sample size		
Haplotype setting	*Odds ratio*	*Frequency*	*400*	*1000*	*2000*
HS1	3	0.005	25	59	89
	2	0.010	32	51	72
	1	—	2.5	4.75	5.25
HS2	3	0.005	28	55	89
	2	0.010	14	30	59
	1	—	4.14	4.29	4.29
HS3	3	0.005	21	47	85
	2	0.010	22	29	64
	1	—	3	3.9	4

Note that frequencies for these individual haplotypes are available in Table 1.
[a]Lines with odds ratios > 1 denote power.
[b]Line with odds ration being 1 denote average type I error rate over all unassociated haplotypes.

Table 4 Potential risk or protective haplotypes for AMD

Haplotype[a]	*Direction*	*Case freq*	*Control freq*	*P-value*
CGGAGCG	Risk	0.5779	0.3151	<0.0001
CGTACTG	Protective	0.0560	0.1477	0.0002
CGTGCCT	Risk	0.0143	0.0000	0.0344
CATACTG	Protective	0.0889	0.1654	0.0018
TGTACTG	Protective	0.0269	0.0861	0.0010
TGTACTT	Risk	0.0111	0.0000	0.0124
TGTGCCG	Protective	0.0000	0.0102	0.0135
TGTGCCT[b]	Protective	0.1306	0.1793	0.0652
TGGAGCG[b]	Risk	0.0844	0.0532	0.0959

[a]Haplotypes in the top segment are found to be significantly associated with AMD by hapKL. In the bottom segment, the two haplotypes that are not significant at the 5% level, but are listed for ease of comparison with results in LBL[6] to illustrate the effect of the selection of a base haplotype.
[b]These two are base haplotype and significant haplotype, respectively in LBL analysis.

haplotype, the effects of other protective haplotypes would be reduced, whereas the effects of risk haplotypes would be intensity. The uncovering of three additional protective haplotypes that match the partial result in Spencer *et al*[11] using hapKL substantiates this theoretical analysis. Conversely, if a risk haplotype is designated as the reference haplotype, then the effects of other risk haplotype would be reduced, while the effects of protective haplotype would be increased. Therefore, a method such as hapKL without the need to subjectively select a reference haplotype is more objective and preferred if the power of the method is comparable to other methods.

There has been a myriad of new statistical methods proposed for detecting association of rare variants with common diseases. These methods may be generally classified as 'collapsing' based and 'haplotyping' based, with the former being applicable to rare SNVs from sequencing data, whereas the latter proposed primarily for using the 'old' GWAS common SNP data. Collapsing-based methods, without a doubt, constitute the predominant class, composed of most of the rare variant association methods proposed to date.[21–25] However, recent works have suggested that haplotyping-based methods may offer advantages and can be more powerful than collapsing methods for some underlying disease settings.[7,8,10] It is in that vein that hapKL is proposed to further explore the benefits of haplotyping-based approaches using existing common SNP data for detecting rare variants that are associated with common diseases without the need to rely on newer sequencing data. The results presented in this paper indicate that hapKL provides an enrichment to the toolkit of haplotype approaches.

Despite the major advantages discussed above, hapKL has its own limitations. First, from a theoretical point of view, the measure hapKL does not have an upper bound, as we mentioned earlier. This undesirable property makes it difficult to assess significance in practice, necessitating the use of a Monte Carlo simulation technique to assess whether a divergence is significantly large than expected under the null hypothesis of no association, as we have done in this paper. This adds to the computational cost, especially if the number of haplotypes is large. For example, for the AMD data, there are 16 haplotypes with nonzero frequencies, and an analysis with 500 Monte Carlo simulations for computing the *P*-values required 79.43 s on a Red Hat Enterprise Linux with a Pentium 4, 3.0 GHz processor. (In comparison, the LBL analysis took 76.13 s.) Given the advantage of nonparametric approaches for detecting rare haplotype–disease association, especially its ability to identify a null haplotype for subsequent analysis, it would be of interest to explore other divergence measures. In particular, a class of divergence measures based on the Shannon entropy, such as the Jensen–Shannon divergence measure, which does have an upper bound,[26] is most promising, as measures with an upper bound may be exploited for the assessment of significance, especially if the bound is tight.

Second, for model-based methods, covariates, such as environmental risk factors, can be accommodated; however, this is not possible in hapKL. The inability of KL for incorporating covariates may lead to false positives if confounders exist. On the other hand, for model-based approaches, if the model or even just the reference haplotype is specified incorrectly, the results can also be erroneous. As such, a potential alternative is to combine these two types of procedures to achieve a better solution. In other words, it would be advisable to run hapKL first to ascertain which haplotypes are unassociated and select the most appropriate one from this pool of candidates as the reference haplotype in a model-based approach to mitigate the effect of an incorrectly selected neutral haplotype. In this regard, one may view hapKL as a preliminary screening tool to aid the selection of a base haplotype before a more elaborate, model-based, analysis that can incorporate environmental and other factors.

CONFLICT OF INTEREST

The author declares no conflict of interest.

ACKNOWLEDGEMENTS

We thank the editor and two anonymous reviewers for their helpful comments and suggestions. This work was partially supported by NSF Grant DMS-1208968. The MAD data set used for the analyses described in this manuscript was obtained from the NEI Study of Age-Related Macular Degeneration (NEI-AMD) Database found at http://www.ncbi.nlm.nih.gov/gap.

1 Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.

2 Li Y, Sung W-K, Liu JJ: Association mapping via regularized regression analysis of single-nucleotide-polymorphism haplotypes in variable-sized sliding windows. *Am J Hum Genet* 2007; **80**: 705–715.

3 Guo W, Lin S: Generalized linear modeling with regularization for detecting common disease rare haplotype association. *Genet Epidemiol* 2009; **33**: 308–316.

4 Tzeng J-Y, Bondell HD: A comprehensive approach to haplotype-specific analysis by penalized likelihood. *Eur J Hum Genet* 2010; **18**: 95–103.

5 Li J, Zhang K, Yi N: A Bayesian hierarchical model for detecting haplotype–haplotype and haplotype–environment interactions in genetic association studies. *Hum Heredity* 2011; **71**: 148–160.

6 Biswas S, Lin S: Logistic Bayesian Lasso for identifying association with rare haplotypes and application to age-related macular degeneration. *Biometrics* 2012; **68**: 587–597.

7 Lin W-Y, Yi N, Zhi D, Zhang K, Gao G, Tiwari HK, Liu N: Haplotypebased methods for detecting uncommon causal variants with common SNPs. *Genet Epidemiol* 2012; **36**: 572–582.

8 Lin W-Y, Yi N, Lou X-Y, Zhi D, Zhang K, Gao G, Tiwari HK, Liu N: Haplotype kernel association test as a powerful method to identify chromosomal regions harboring uncommon causal variants. *Genet Epidemiol* 2013; **37**: 560–570.

9 Biswas S, Xia S, Lin S: Detecting rare haplotype-environment interaction with logistic Bayesian Lasso. *Genet Epidemiol* 2014; **38**: 31–41.

10 Wang M, Lin S: famLBL: detecting rare haplotype disease association based on common SNPs using case–parent triads. *Bioinformatics* 2014; **30**: 2611–2618.

11 Spencer KL, Hauser MA, Olson LM, Schnetz-Boutaud N, Scott WK, Schmidt S, Gallins P, Agarwal A, Postel EA, Pericak-Vance MA, Haines JL: Haplotypes spanning the complement factor H gene are protective against age-related macular degeneration. *Invest Ophthalmol Vis Sci* 2007; **48**: 4277–4283.

12 Schaid D, Rowland C, Tines D, Jacobson R, Poland G: Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 2002; **70**: 425–434.

13 Lake SL, Tantisira K, Sulverman E, Weiss S, Laird N, DJ S: Estimation and testing of haplotype-environment interactions when linkage phase is ambiguous. *Hum Heredity* 2003; **55**: 56–65.

14 Burkett K, Graham J, McNeney B: Hapassoc: Software for likelihood inference of trait associations with SNP haplotypes and other attributes. *J Statist Softw* 2006; **16**: 1–19.

15 Tibshirani R: Regression shrinkage and selection via the Lasso. *J R Statist Soc Ser B* 1996; **58**: 267–288.

16 Zhu X, Feng T, Li Y, Lu Q, Elston RC: Detecting rare variants for complex traits using family and unrelated data. *Genet Epidemiol* 2010; **34**: 171–187.

17 Kullback S, Leibler R: On information and sufficiency. *Ann Math Statist* 1951; **22**: 79–86.

18 Vajda I: Note on discrimination information and variation. *IEEE Trans Inform Theory* 1970; **16**: 771–773.

19 Klein R, Zeiss C, Chew E, Tsai J, Sackler R, Haynes C, Henning A, San-Giovanni J, Mane S, Mayne S, Bracken M, Ferris F, Ott J, Barnstable C, Hoh J: Complement factor H polymorphism in age-related macular degeneration. *Science* 2005; **308**: 385–389.

20 Li M, Atmaca-Sonmez P, Othman M, Branham KEH, Khanna R, Wade MS, Li Y, Liang L, Zareparsi S, Swaroop A, Abecasis GR: *CFH* haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. *Nat Genet* 2006; **38**: 1049–1054.

21 Dering C, Hemmelmann C, Pugh E, Ziegler A: Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genet Epidemiol* 2011; **35**: S12–S17.

22 Li B, Leal SM: Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am J Hum Genet* 2008; **83**: 311–321.

23 Madsen BE, Browning SR: A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009; **5**: e1000384.

24 Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ: Testing for an unusual distribution of rare variants. *PLoS Genet* 2011; **7**: e1001322.

25 Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011; **89**: 82–93.

26 Lin J: Divergence measures based on the shannon entropy. *IEEE Trans Inform Theory* 1991; **37**: 145–151.