Journal of Human Genetics (2012) 57, 734–737

www.nature.com/jhg

ORIGINAL ARTICLE

Genotype-based association analysis via entropy

Yu-Mei Li and Yang Xiang

With the advent of large-scale genotyping technologies, enormous quantities of genotype data that were generated have been well exploited through phased haplotypes, and the haplotype-based association study is used as one of the major statistical methods for gene mapping of human complex traits. However, haplotype-based method depends on the information of haplotype frequencies, which results in infeasible computation when haplotypes are not directly observed. This paper provides a genotype-based statistic with multiple tightly linked markers for association analysis using entropy theory. The statistic here does not require haplotype phasing and only requires genotype data. The distribution and the power of the statistic are investigated by simulative study. The results show that the statistic has very reasonable performance. We demonstrated the powerfulness of the statistic by applying our approach to a specific example on hereditary hemochromatosis.
Journal of Human Genetics (2012) **57,** 734–737; doi:10.1038/jhg.2012.102; published online 23 August 2012

Keywords: association analysis; entropy; linkage disequilibrium

INTRODUCTION

Linkage disequilibrium (LD), the non-random co-occurrence of alleles from different loci, has a fundamental role in genetic studies as a tool for gene mapping of human complex traits. However, the level of LD is often influenced by a number of factors, including genetic linkage, selection, the rate of recombination and mutation, genetic drift, non-random mating, population structure and other non-biological forces. These bring a great many challenges for LD mapping or association analysis in genetic studies. One of those challenges is to develop novel statistical methods to improve the power of gene mapping. Although many LD methods have been well developed currently for complex disease genes, haplotype-based analysis is one of the major statistical methods because haplotypes of multiple single-nucleotide polymorphisms (SNPs) are considered a more informative format of polymorphisms for genetic analysis than single SNP.[1]

The classical haplotype-based statistic is to compare haplotype frequencies between affected and unaffected individuals[1,2] or to compare haplotype similarities between affected and unaffected individuals.[3,4] Recently, Zhao *et al.*[5] proposed an entropy-based statistic T_{PE} for a genome-wide association study through a non-linear transformation of haplotype frequencies. Nonetheless, the results of these methods are not uniformly consistent.[4,5] An important reason is that haplotype-based method depends on the information of haplotype frequencies, which results in infeasible computation for estimating haplotype frequencies when haplotypes are not directly observed.

In this paper, we will propose an entropy-based statistic; here we denote it as T_{GE}, as an alternative method of Zhao *et al.*,[5] which only allows for genotype data at linked markers. We will investigate the performance of the statistic T_{GE} by computer simulations and apply it to a real data set on hereditary hemochromatosis (HH).

MATERIALS AND METHODS

Composite LD measure

We consider two marker loci '1' and '2', with co-dominant alleles A and a for loci 1 and B and b for loci 2. The frequencies of alleles A, a, B and b are given by P_A, P_a, P_B and P_b, and the frequencies of genotype AA, Aa, aa, BB, Bb and bb are given by P_{AA}, P_{Aa}, P_{aa}, P_{BB}, P_{Bb} and P_{bb}, respectively. Let P_{AB}, P_{Ab}, P_{aB} and P_{ab} be the haplotype frequencies of AB, Ab, aB and ab, respectively. Let $P_{/}$ be the joint frequency of alleles for loci 1 and 2 in two different gametes, for example, $P_{A/B}$ is the joint frequency of alleles A and B in two different gametes. The composite LD coefficient is defined as $\Delta_{AB} = P_{AB} + P_{A/B} - 2P_AP_B$, and an estimator of Δ_{AB} is defined as $\hat{\Delta}_{AB} = \frac{N_{AB}}{N} - 2\hat{P}_A\hat{P}_B$, where $N_{AB} = 2n_{AABB} + n_{AABb} + n_{AbBB} + 1/2n_{AaBb}$, N is the total number of subjects, n is a count of the number of subjects with the phenotype indicated by its subscript, and $\hat{P}_A$ and $\hat{P}_B$ are estimates of allele frequencies. It can be seen that these estimators only rely on the information of genotype. We define a two-dimensional random variable $X = (X_1, X_2)^T$ as the state of two alleles A and B, where X_1 and X_2 denote the number of copies 'A' and 'B' for marker '1' and '2', respectively. The probability distributions of X_1 and X_2 are described in Table 1. It can be shown that $\mathrm{var}(X_1) = P_{Aa} + 4P_{AA} - 4P_A^2$ and $\mathrm{var}(X_2) = P_{Bb} + 4P_{BB} - 4P_B^2$, and the covariance of X_1 and X_2 is equal to twice that of the composite LD coefficient Δ_{AB}:$\mathrm{cov}(X_1, X_2) = 2\Delta_{AB}$.

The entropy-based statistic T_{GE}

Suppose that there are k markers, each of which has two alleles A and a. The frequency of allele A is denoted by P_i for ith marker ($i = 1, \ldots, k$). The entropy of the frequencies P_i for the ith marker is defined as $H_i = -P_i \log P_i$.[6] Denote the first partial derivatives of the entropy H_i with respect to the frequency P_j as z_{ij}: $z_{ij} = -1 - \log P_i$ for $i = j$, $z_{ij} = 0$ for $i \neq j$, $i, j = 1, \ldots, k$. To simplify our presentation, a measure with a superscript 'A' indicates a measure in affected

Department of Mathematics, Huaihua University, Huaihua, China
Correspondence: Dr Y-M Li, Department of Mathematics, Huaihua University, Huaihua, Hunan 418008, China.
E-mail: liym74@yahoo.com or lymmail@126.com
Received 7 June 2012; revised 16 July 2012; accepted 20 July 2012; published online 23 August 2012

Table 1 The probability distributions of X_1 and X_2

X_1,X_2	0	1	2
P_{X_1}	P_{aa}	P_{Aa}	P_{AA}
P_{X_2}	P_{bb}	P_{Bb}	P_{BB}

individuals and a measure with a superscript 'C' indicates a measure in unaffected individuals. The frequencies of alleles A in affected individuals for k markers are given by the vector $P^{\mathrm{A}}=(P_1^{\mathrm{A}},P_2^{\mathrm{A}},\ldots,P_k^{\mathrm{A}})^{\mathrm{T}}$ and the frequencies of alleles A in unaffected individuals are given by the vector $P^{\mathrm{C}}=(P_1^{\mathrm{C}},P_2^{\mathrm{C}},\ldots,P_k^{\mathrm{C}})^{\mathrm{T}}$. Let $H^{\mathrm{A}}=(H_1^{\mathrm{A}},H_2^{\mathrm{A}},\ldots,H_k^{\mathrm{A}})^{\mathrm{T}}$ and $H^{\mathrm{C}}=(H_1^{\mathrm{C}},H_2^{\mathrm{C}},\ldots,H_k^{\mathrm{C}})^{\mathrm{T}}$. Denote X_i as the state of allele A, that is, the number of copies 'A' for marker 'i', $i=1,\ldots,k$. Define a k-dimensional random variable $X=(X_1,X_2,\ldots X_k)^{\mathrm{T}}$. Suppose that there are n^{A} affected individuals and n^{C} unaffected individuals ($n^{\mathrm{A}}+n^{\mathrm{C}}=2N$). The total number of the alleles A for marker 'i' in n^{A} affected individuals and n^{C} unaffected individuals are denoted as N_i^{A} and N_i^{C}, respectively. Let $X_i^{\mathrm{A}}=(X_{1i}^{\mathrm{A}},X_{2i}^{\mathrm{A}},\ldots,X_{ki}^{\mathrm{A}})^{\mathrm{T}}$ and $X_j^{\mathrm{C}}=(X_{1j}^{\mathrm{C}},X_{2j}^{\mathrm{C}},\ldots,X_{kj}^{\mathrm{C}})^{\mathrm{T}}$ be the state of allele A for the ith ($i=1,\cdot 2,\ldots,\cdot n^{\mathrm{A}}$) affected individual and the jth ($j=1,\cdot 2,\ldots,\cdot n^{\mathrm{C}}$) unaffected individual, respectively. It is easy to see that

$$\bar{X}^{\mathrm{A}}=\tfrac{1}{n^{\mathrm{A}}}\left(\sum_{i=1}^{n^{\mathrm{A}}}X_{1i}^{\mathrm{A}},\sum_{i=1}^{n^{\mathrm{A}}}X_{2i}^{\mathrm{A}},\ldots,\sum_{i=1}^{n^{\mathrm{A}}}X_{ki}^{\mathrm{A}}\right)^{\mathrm{T}}=\tfrac{1}{n^{\mathrm{A}}}(N_1^{\mathrm{A}},N_2^{\mathrm{A}},\ldots,N_k^{\mathrm{A}})^{\mathrm{T}},$$

and

$$\bar{X}^{\mathrm{C}}=\tfrac{1}{n^{\mathrm{C}}}\left(\sum_{i=1}^{n^{\mathrm{C}}}X_{1i}^{\mathrm{C}},\sum_{i=1}^{n^{\mathrm{C}}}X_{2i}^{\mathrm{C}},\ldots,\sum_{i=1}^{n^{\mathrm{C}}}X_{ki}^{\mathrm{C}}\right)^{\mathrm{T}}=\tfrac{1}{n^{\mathrm{C}}}(N_1^{\mathrm{C}},N_2^{\mathrm{C}},\ldots,N_k^{\mathrm{C}})^{\mathrm{T}}$$

Let $N^{\mathrm{A}}=(N_1^{\mathrm{A}},N_2^{\mathrm{A}},\ldots,N_k^{\mathrm{A}})^{\mathrm{T}}$ and $N^{\mathrm{C}}=(N_1^{\mathrm{C}},N_2^{\mathrm{C}},\ldots,N_k^{\mathrm{C}})^{\mathrm{T}}$. Note that $N^{\mathrm{A}}/2n^{\mathrm{A}}$ and $N^{\mathrm{C}}/2n^{\mathrm{C}}$ are the maximum likelihood estimator of the vector P^{A} and P^{C}, respectively, here; let $\hat{P}^{\mathrm{A}}=N^{\mathrm{A}}/2n^{\mathrm{A}}$ and $\hat{P}^{\mathrm{C}}=N^{\mathrm{C}}/2n^{\mathrm{C}}$, and from the asymptotic normality of the maximum likelihood estimator, we have

$$\sqrt{n^{\mathrm{A}}}\left(\hat{P}^{\mathrm{A}}-P^{\mathrm{A}}\right)\to\left(0,\frac{\sum^{\mathrm{A}}}{4}\right),\ \sqrt{n^{\mathrm{C}}}\left(\hat{P}^{\mathrm{C}}-P^{\mathrm{C}}\right)\to\left(0,\frac{\sum^{\mathrm{C}}}{4}\right),$$

where $\sum^{\mathrm{A}}=(\sigma_{ij}^{\mathrm{A}})_{k\times k}$, $\sigma_{ii}^{\mathrm{A}}=\mathrm{var}(X_i^{\mathrm{A}})$, $\sigma_{ij}^{\mathrm{A}}=\mathrm{cov}(X_i^{\mathrm{A}},X_j^{\mathrm{A}})=2\Delta_{ij}^{\mathrm{A}}$ for $i\neq j$, and $\sum^{\mathrm{C}}=(\sigma_{ij}^{\mathrm{C}})_{k\times k}$, $\sigma_{ii}^{\mathrm{C}}=\mathrm{var}(X_i^{\mathrm{C}})$, $\sigma_{ij}^{\mathrm{C}}=\mathrm{cov}(X_i^{\mathrm{C}},X_j^{\mathrm{C}})=2\Delta_{ij}^{\mathrm{C}}$ for $i\neq j$,[7] here, σ_{ii}^{A} is the variance of X_i for marker 'i', and Δ_{ij}^{A} and Δ_{ij}^{C} are the composite LD coefficient for marker 'i' and marker 'j', respectively.

Let $Z^{\mathrm{A}}=(z_{ij}^{\mathrm{A}})_{k\times k}$,$Z^{\mathrm{C}}=(z_{ij}^{\mathrm{C}})_{k\times k}$, $W^{\mathrm{A}}=Z^{\mathrm{A}}\Sigma^{\mathrm{A}}(Z^{\mathrm{A}})^{\mathrm{T}}$ and $W^{\mathrm{C}}=Z^{\mathrm{C}}\Sigma^{\mathrm{C}}(Z^{\mathrm{C}})^{\mathrm{T}}$. Let $\hat{H}^{\mathrm{A}}$, $\hat{H}^{\mathrm{C}}$, $\hat{W}^{\mathrm{A}}$ and $\hat{W}^{\mathrm{C}}$ be the estimators of H^{A}, H^{C}, W^{A} and W^{C}, respectively. Then the entropy-based statistic can be defined as

$$T_{\mathrm{GE}}=4(\hat{H}^{\mathrm{A}}-\hat{H}^{\mathrm{C}})^{\mathrm{T}}\left(\frac{\hat{W}^{\mathrm{A}}}{n^{\mathrm{A}}}+\frac{\hat{W}^{\mathrm{C}}}{n^{\mathrm{C}}}\right)^{-1}(\hat{H}^{\mathrm{A}}-\hat{H}^{\mathrm{C}})$$

The statistic T_{GE} is asymptotically a central $\chi^2_{(k)}$ distribution under the null hypothesis of no association and is asymptotically a non-central $\chi^2_{(k)}$ distribution[8] with the non-centrality parameter $\lambda_{\mathrm{GE}}=4(H^{\mathrm{A}}-H^{\mathrm{C}})^{\mathrm{T}}(\frac{W^{\mathrm{A}}}{n^{\mathrm{A}}}+\frac{W^{\mathrm{C}}}{n^{\mathrm{C}}})^{-1}(H^{\mathrm{A}}-H^{\mathrm{C}})$ under the alternative hypothesis of association. When the inverse of the matrix $\frac{W^{\mathrm{A}}}{n^{\mathrm{A}}}+\frac{W^{\mathrm{C}}}{n^{\mathrm{C}}}$ does not exist, the generalized inverse of the matrix will be used.

RESULTS

Distribution of the statistic T_{GE} and type I error rate

To assess the distribution of the T_{GE} in finite samples, a simulation study is performed to investigate the distribution of the statistic T_{GE} under the null hypothesis of no association. The simulations using the computer program SNaP[9] are implemented similar to those in Zhao *et al.*[5] We consider two biallelic marker loci that generate four haplotypes (*AB*, *Ab*, *aB* and *ab*) with frequencies 0.2952, 0.2562, 0.1957 and 0.2529, respectively. We randomly generate 20 000 individuals in the general population and divide them into equal groups of cases and controls. Then, we randomly sample N individuals (here, $n^{\mathrm{A}}=n^{\mathrm{C}}=N$) each from the cases and controls with 10 000 replicates; and thus, it follows 10 000 values of the statistic T_{GE}. Figure 1 plots the histograms of the test statistic T_{GE} when $2N$ is 200. It can be seen that the distributions of T_{GE} is similar to the theoretical central $\chi^2_{(2)}$ distributions. For a given significance level α ($\alpha=0.05$), the type I error rate is estimated as the proportion of rejecting the null hypothesis in 10 000 replicates performed when the null hypothesis holds. The estimated type I error rates for sample sizes ($2N$) from 200 to 1000 individuals are exhibited in Table 2. We can see that the type I error rates are around the nominal levels $\alpha=0.05$. All these indicate the validity of the statistic T_{GE}.

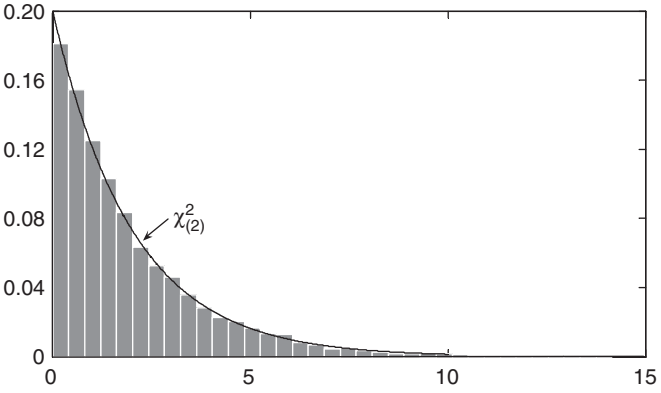

Figure 1 Distribution of the test statistic T_{GE} using genotypes of two markers indicates $\chi^2_{(2)}$ distribution.

Table 2 Estimated type I error rates of the statistic T_{GE} for 10 000 simulations

	Estimated type I error rate		
Sample size	$\alpha=0.05$	$\alpha=0.01$	$\alpha=0.001$
200	0.0478	0.0082	0.0014
400	0.0502	0.0102	0.0010
600	0.0470	0.0091	0.0008
800	0.0495	0.0090	0.0012
1000	0.0491	0.0093	0.0009

The power of the test statistic T_{GE}

To evaluate the power of the statistic T_{GE} and compare the power of the T_{GE} with that of the statistic T_{PE}, we performed simulations under the alternative hypothesis of association. Consider a biallelic disease locus with alleles D and d and two biallelic marker loci that generate four haplotypes (*AB*, *Ab*, *aB* and *ab*) with frequencies 0.2952, 0.2562, 0.1957 and 0.2529, respectively. The disease allele frequency is taken as $P_{\mathrm{D}}=0.3$. The penetrances of genotypes *DD*, *Dd* and *dd* are denoted by f_{11}, f_{12} and f_{22}, respectively. The overall LD[10] between the four haplotypes and the allele D at the disease locus are chosen as 0.0728, -0.0448, 0.0192 and -0.0472, respectively. We consider four genetic models: additive model, dominant model, recessive model and multiplicative model. Under a specific genetic model, for a particular sample size, the power is estimated as the proportion of rejecting the null hypothesis in 10 000 replicates performed when the alternative hypothesis holds. Figure 2 shows the power curves against sample

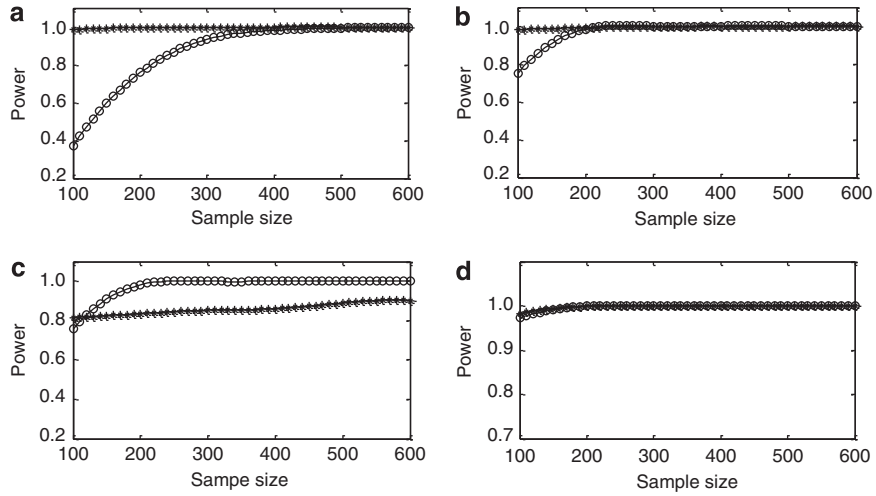

Figure 2 Power curves against sample size at the 0.05 significance level for the statistic T_{GE} (o) and T_{PE} (*) under the recessive model (**a**, $f_{11}=1$, $f_{12}=0.1$, $f_{22}=0.1$), the additive model (**b**, $f_{11}=1$, $f_{12}=0.5$, $f_{22}=0$), the dominant model (**c**, $f_{11}=1$, $f_{12}=1$, $f_{22}=0$) and the multiplicative model (**d**, $f_{11}=0.81$, $f_{12}=0.045$, $f_{22}=0.0025$).

size at the 0.05 significance level for T_{GE} and T_{PE}. It can be seen that when the sample size is larger than 100, the statistic T_{GE} has consistently higher power than the statistic T_{PE} under the dominant model. The power of the T_{GE} is almost the same as that of the T_{PE} under the multiplicative model and the additive model, except when the sample size is smaller than 200 for the additive model. But we can see that the statistic T_{PE} dominates the statistic T_{GE} under the recessive model when the sample size is smaller than 400.

Applications to HH

HH is inherited as a recessive disease resulting in excessive iron absorption from diet and leads to chronic disease and early death. It is one of the most common inherited diseases among people of European descent, with an estimated prevalence of 1/200 to 1/400, with an even higher prevalence likely in the Irish population.[11] About 1 in 8 to 1 in 10 Australians of Northern European ancestry are genetic carriers for HH.[12] In the course of cloning the hemochromatosis gene, genotypes in 101 HH cases and 64 controls were genotyped at 43 microsatellite repeat markers that span the 6.5-Mb HH gene region[13] (http://link.springer.de/link/service/journals/00439/index.htm). We analyzed the data using the statistic T_{GE}. To simplify our computation, four markers are used in our analysis: 2229, 2241, 2242 and 2236. We take the ancestral allele given in Thomas *et al.*[13] at each marker as allele *A* and all the other alleles as allele *a*. In Table 3, we present the values of the statistic T_{GE} and the corresponding *P*-values for testing the association of two, three and four SNP markers with HH. It is evident that the statistic T_{GE} obtained a very small *P*-value.

Table 3 Tests of association between hemochromatosis genotype and HH

	The values of T_{GE} (P-*value*)
Two markers	
D6S2229, D6S2242	99.33 ($<10^{-16}$)
D6S2229, D6S2241	113.80 ($<10^{-16}$)
D6S2229, D6S2236	63.29 (1.79×10^{-14})
D6S2242, D6S2241	95.65 ($<10^{-16}$)
D6S2242, D6S2236	67.78 (1.88×10^{-15})
D6S2241, D6S2236	64.75 (8.65×10^{-15})
Three markers	
D6S2229, D6S2242, D6S2241	133.19 ($<10^{-16}$)
D6S2229,D6S2242, D6S2236	109.02 ($<10^{-16}$)
D6S2229, D6S2241, D6S2236	121.45 ($<10^{-16}$)
D6S2242, D6S2241, D6S2236	104.78 ($<10^{-16}$)
Four markers	
D6S2229, D6S2242, D6S2241, D6S2236	146.14 ($<10^{-16}$)

DISCUSSION

With the availability of high-density maps of SNP markers, population-based LD mapping or association study provides an unprecedented opportunity for identifying genetic variants that influence human complex trait. Haplotype-based analysis is used as one of the major statistical methods for mapping gene of human complex trait. When only genotype data at multiple loci are collected from a sample of unrelated individuals, haplotype-based methods need estimating haplotype phases and frequencies. Although current computational and laboratory methods promise improved determination of haplotype phase, the haplotype-based method is not yet cost-effective.[14,15] Here, we proposed an entropy-based statistic using genotype data for association analysis. The validity and the power of the statistic are demonstrated by simulation analysis and case study. In contrast with haplotype-based association study, our method does not need the haplotype information, and the power of our method is sometimes higher than that of haplotype-based method (for example, the method of Zhao *et al.*[5]) when the sample size is not very small. Our paper here mainly investigated the effect of mode of inheritance on the power and the effect of LD between SNPs on the power is not examined. More work awaits to be done with the effect of LD between SNPs in the future.

ACKNOWLEDGEMENTS

This study was supported by the Foundation of Hunan Educational Committee (11B095) and National Social Science Fund Youth Project (11CTJ003).

1 Akey, J., Jin, L. & Xiong, M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur. J. Hum. Genet.* **9,** 291–300 (2001).
2 Chapman, N. H. & Wijsman, E. M. Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. *Am. J. Hum. Genet.* **63,** 1872–1885 (1998).
3 Bourgain, C., Genin, E., Margaritte-Jeannin, P. & Clerget-Darpoux, F. Maximum identity length contrast: a powerful method for susceptibility gene detection in isolated populations. *Genet. Epidemiol. Suppl.* **21,** S560–S564 (2001).
4 Tzeng, J. Y., Devlin, B., Wasserman, L. & Roeder, K. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am. J. Hum. Genet.* **72,** 891–902 (2003).
5 Zhao, J. Y., Boerwinkle, E. & Xiong, M. An entropy-based statistic for genomewide association studies. *Am. J. Hum. Genet.* **77,** 27–40 (2005).
6 Shannon, C. E. A mathematical theory of communication. *Bell. Systems Tech. J.* **27,** 379–423 (1948).
7 Rao, C. R. *Linear Statistical Inference and its Applications* (John Wiley, New York, 1973).
8 Lehmann, E. L. *Theory of Point Estimation* (John Wiley & Sons, New York, 1983).
9 Nothnagel, M. Simulation of LD block-structured SNP haplotype data and its use for the analysis of case-control data by supervised learning methods. *Am. J. Hum. Genet.* **71**(Suppl), A2363 (2002).
10 Xiong, M., Zhao, J. Y. & Boerwinkle, E. Haplotype block linkage disequilibrium mapping. *Front. Biosci.* **8,** 85–93 (2003).
11 Feder, J. N., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D. A., Basava, A. A. *et al.* A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat. Genet.* **13,** 399–408 (1996).
12 Edward, C. Q., Griffen, L. M., Goldgar, D., Drummond, C., Solnick, M. H. & Kushner, J. P. Prevalence of haemochromatosis among 11,065 presumbly healthy donors. *N Engl. J. Med.* **318,** 1355–1362 (1988).
13 Thomas, W., Fullan, A., Loeb, D. B., McClelland, E. E., Bacon, B. R. & Wolff, R. K. A haplotype and linkage disequilibrium analysis of the hereditary haemochromatosis gene region. *Hum. Genet.* **102,** 517–525 (1998).
14 Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12,** 703–714 (2011).
15 Ruiz-Marín, M., Matilla-García, M., Cordoba, J. A., Susillo-González, J. L., Romo-Astorga, A., González-Pérez, A. *et al.* An entropy test for single-locus genetic association analysis. *BMC Genet.* **23,** 11–19 (2010).