European Journal of Human Genetics (2011) 19, 801–806

www.nature.com/ejhg

ARTICLE

Simple strategies for haplotype analysis of the X chromosome with application to age-related macular degeneration

Renfang Jiang[1], Jianping Dong[1], Jungnam Joo[2], Nancy L Geller[3] and Gang Zheng*[,3]

For haplotype analysis of the X chromosome, haplotype-sharing (HS) statistics with sliding windows are defined for males and females separately, which are then combined to a single HS test for the X chromosome. When independent replication samples are not available, the training-testing sets approach is used to validate this procedure and a permutation method is used to obtain its *P*-value. We applied this method to the X chromosome (with 1804 SNPs) for age-related macular degeneration (AMD). We found a window of five SNPs over a 272 kb region associated with AMD after Bonferroni correction. An examination of the odds ratio and the population attributable risks revealed a disease-preventive haplotype, *ATGAC*, on these five SNPs. For elderly females without this haplotype, the likelihood of AMD is increased by a factor of 4.75 with a 95% confidence interval (1.43, 15.82). The frequency of *ATGAC* in HapMap CEU is 0.276. These five SNPs are covered by the gene *DIAPH2*, which is known to cause premature ovarian failure (POF) in females. Our results indicated that *DIAPH2* may be a polygenic pleiotropy for POF and AMD.

European Journal of Human Genetics (2011) **19,** 801–806; doi:10.1038/ejhg.2011.35; published online 9 March 2011

Keywords: case–control studies; GWAS; haplotype analysis; sex-linked genes

INTRODUCTION

Statistical methods for the analysis of autosomal chromosomes cannot be directly applied to the X chromosome, because males and females need to be treated separately. Single-marker analysis for the X chromosome has been considered.[1,2] It is well known that haplotype analysis is often more powerful than single-marker analysis. We propose a simple test for haplotype analysis of the X chromosome. It applies haplotype-sharing (HS) statistics for males and females separately and combines them as a single HS test for the X chromosome. It incorporates sliding-window approach and a permutation procedure to obtain the *P*-value.

The HS test is applied to the X chromosome of the genome-wide association study (GWAS) of National Eye Institute: Age-Related Eye Disease Study (NEI-AREDS) using 1804 SNPs with two analysis strategies.

METHODS

Haplotype-sharing (HS) association test

The HS tests, developed for autosomal regions, are adopted here to analyze the X chromosome. Let h_1 and h_2 be two haplotypes on a chromosome region. Their sharing length around a particular marker l is the length of the contiguous chromosomal region on which the two haplotypes are identical by states (IBS). Two haplotypes are compared at each marker starting from marker l. If two haplotypes are IBS at markers l_1, l_1+1,…, l,…, l_2−1, l_2, and they are not IBS at markers l_1-1and l_2+1, then the distance between markers l_1 and l_2, denoted as $S_{h_1,h_2}(l) = l_2 - l_1$, is the sharing length of the two haplotypes around marker l.[3] For a given haplotype h_i, its sharing score around marker l is $N^{-1}\sum_{j=1}^{N} S_{h_i,h_j}(l)$, where N is the number of haplotypes in the data.

For a region with k SNPs, suppose that there are n haplotypes from cases and m haplotypes from controls and $N=n+m$. Let $x_i(l) = N^{-1}\sum_{j=1}^{N} S_{h_i,h_j}(l)$, $i=1,\ldots, n$ be the haplotype-sharing score of haplotypes from cases, and $y_j(l) = N^{-1}\sum_{i=1}^{N} S_{h_i,h_j}(l)$, $j=1\ldots, m$ from controls. Let $\bar{x}(l) = \sum_i x_i(l)/n$, $\bar{y}(l) = \sum_j y_j(l)/m$, and $s_p(l)$ be the pooled SD given by $s_p^2(l) = \{(n-1)s_1^2(l) + (m-1)s_2^2(l)\}/(n+m-2)$, where $s_1^2(l) = \sum_i \{x_i(l) - \bar{x}(l)\}^2/(n-1)$ and $s_2^2(l) = \sum_j \{y_j(l) - \bar{y}(l)\}^2/(m-1)$. For large samples, $T(l) = \{\bar{x}(l) - \bar{y}(l)\}/\{s_p(l)\}^2$ follows, approximately, a χ^2-distribution with one degree of freedom. Note that $T(l)$ is a statistic around marker l. The HS test for a region containing k SNPs is $T_k = \max_{1\le l\le k} T(l)$. This HS test is a special case of Tzeng *et al*[3] and can also be derived from Zhang *et al*.[4]

The above HS test assumes that phase information is known. This is true for males on the X chromosome, but generally not true for females. Suppose all females are unrelated individuals from a population with Hardy–Weinberg equilibrium (HWE). The maximum-likelihood estimates of haplotype frequencies can be obtained by the expectation-maximization (EM) algorithm.[5] Then, we apply HS tests to males and females separately, denoted as $T_{k,\text{male}}$ and $T_{k,\text{female}}$ and combined them as $T_k = T_{k,\text{male}} + T_{k,\text{female}}$. For comparison, T_k is also calculated for the pooled males and females, denoted as $T_{k,\text{pooled}}$. However, in $T_{k,\text{pooled}}$, the phase information in males is not used to infer that in females.

Sliding window for HS association test

To carry out the HS test, we consider a sliding-window framework. In a large-scale haplotype analysis, one needs to determine the number of adjacent SNPs considered in the HS test. Presumably, M SNPs are genotyped on a

[1]Department of Mathematical Sciences, Michigan Technological University, Houghton, MI, USA; [2]Cancer Biostatistics Branch, National Cancer Center, Geonggi-do, Korea; [3]Office of Biostatistics Research, National Heart, Lung and Blood Institute, Bethesda, MD, USA
*Correspondence: Dr G Zheng, Office of Biostatistics Research, National Heart, Lung and Blood Institute, 6701 Rockledge Drive, MSC 7913, Bethesda, MD 20892-7913, USA. Tel: +1 301 4351287; Fax: +1 301 4801862; E-mail: zhengg@nhlbi.nih.gov
Received 30 August 2010; revised 28 January 2011; accepted 3 February 2011; published online 9 March 2011

chromosome, where M is much larger than the window size. One should consider a small subset of M SNPs, say k SNPs, which form a window, and the HS test T_k is performed within the window. If a region consists of more than k SNPs, one can slide the window across the region to obtain a sequence of HS tests (for every k contiguous SNPs). Then the maximum of this sequence of HS tests is used to test association for this region, which is also denoted as T_k.

An important issue of this approach is to decide the number of adjacent SNPs k in a window. One approach is to define a window by a linkage disequilibrium (LD) block.[6] As there is only a few distinct haplotypes inside a block because of the strong LD, if the degrees of freedom of a test statistic is directly related to the number of distinct haplotypes, defining the window inside a LD block means a smaller number of distinct haplotypes and that, in turn, results in a smaller number of degrees of freedom and a higher power of the test. The HS test statistic, however, is based on the sharing length of haplotypes. Its degree of freedom does not depend on the number of distinct haplotypes. Thus, defining a window inside a LD block has no advantage. If a window is inside a block, sharing scores of cases and controls could both be large, leading to less or no power for the HS tests. Moreover, haplotypes from cases have common ancestors and if there is no LD among SNPs, their sharing length will decrease rapidly in a few generations because of recombination. At the same time, the sharing length of haplotypes from controls is also small so the HS test cannot detect the difference. Therefore, the SNPs in a window should not be too far apart. Another aspect of the HS test is that the windows should have a minimum number of two SNPs to achieve reasonable power. We used the uniform window size approach.[7–9]

To ensure that the SNPs in a window are still linked, we required the distance between the two adjacent SNPs in a window be <200 kb. If we choose maximum distances between SNPs within chromosome regions smaller than 200 kb, there might be more regions containing a single SNP, which will be excluded from the haplotype analysis.

Two analysis strategies

A simple method for haplotype analysis is the training-testing sets approach, which randomly divides the data into two independent parts: one serves as a training set, and the other, a testing set. We can use this training set for single-marker analysis at each of the SNPs on the X chromosome. Then, we choose tagging SNPs with the largest statistics or smallest P-values from the single-marker analysis. Finally, we apply the HS tests to the testing set on the regions formed by the tagging SNPs. In this approach, the haplotype analysis using the testing set can serve as validation.

Alternatively, in the analysis of GWAS, it is common to conduct single-marker analysis first. Haplotype analysis can be done only on the regions where single-marker analysis is significant after Bonferroni correction or the regions where the top-ranked markers on the basis of single-marker analysis are located. In this case, all samples are used in both single-marker analysis and haplotype analysis. The results can be confirmed by subsequent replication studies using independent samples. We call this approach a two-stage analysis procedure.

We prefer the two-stage analysis approach, when independent replication samples are available. The advantage of this approach is that we do not require the single-marker analysis to be significant, for example, selection of SNPs or regions can be based on the ranks of P-values. If independent replication samples are not available, however, the training-testing sets approach is more appropriate. If there is no replication study in the second approach, false-positive rate of haplotype analysis is much higher because there is high correlation between the single-marker and haplotype analyses under the null hypothesis unless the single-marker and haplotype analyses cover the same region and are both significant after their respective Bonferroni corrections. In Online Supplementary material, we present results from a simple simulation study, which demonstrates that the overall type I error rate is controlled in haplotype analysis with Bonferroni correction only for the number of haplotype analyses, while the SNP in the same region in single-marker analysis is also significant after Bonferroni correction for the total number of SNPs tested.

RESULTS

Data

AMD is the most common cause of severe vision loss. The prevalence of the disorder increases with age, so it has a great impact on the quality of life for the elderly. It is estimated that there are currently about 1.75 million AMD patients in the United States, and is expected to increase to about 3 million in the year 2020.[10] Results of population- and family-based linkage and association studies for AMD have been reported in the literature. For example, a case–control GWAS with 100 000 SNPs was analyzed.[11] This GWAS dataset was a subset of NEI AREDS, in which more extreme phenotypes were selected for cases and controls to enhance the power to detect true associations.[11]

Using the same subset, Zheng *et al*[1] applied quality control to the 2334 SNPs on the X chromosome and conducted single-marker analysis on the 1804 SNPs. The data contained 96 cases (54 females) and 50 controls (23 females) balanced for smoking status, with controls selected to be older than cases. All subjects were self-reported Whites. Population stratification was not found in single-marker analysis. These quality control procedures were similar to those used in GWAS. Among 2334 SNPs on the X chromosome, SNPs due to heterozygosity in males (60 SNPs), monomorphism in either males or females (453 SNPs), with call rate $<75\%$ (4 SNPs), and whose minor allele frequency (MAF) $<1\%$ (13 SNPs) were not analyzed in the single-marker analysis for the X chromosome. Here, we conduct haplotype analysis based on the same 1804 SNPs on the X chromosome.

HS association tests for the AMD data

Although our main analysis is to use the training-testing sets approach, we first test single marker followed by haplotype analysis using all the samples. Zheng *et al*[1] studied six single-marker tests for the X chromosome. We calculated the minimum of P-values of the six statistics and ranked the 1804 SNPs by their minimum P-values.

The 'top SNPs' refer to those with smallest minimum P-values. The top 20% of the SNPs on the X chromosome (360 SNPs) were used for the haplotype analysis using HS tests, although only one of 360 SNPs (rs10521496) was significant after Bonferroni correction, which was also reported before.[1]

The distances between some adjacent SNPs on the X chromosome are quite large, which naturally form chromosomal regions. We defined regions so that the distance between any adjacent SNPs is <200 kb. On the basis of these 360 SNPs, we obtained 77 chromosomal regions after dropping regions with a single SNP. A sliding-window procedure was applied to each region. If a chromosomal region consists of at least five SNPs (or four SNPs, if the maximal window size is four instead of five), we used a uniform window size of five SNPs (or four SNPs); otherwise, the window size is equal to the number of SNPs in the region. In Figure 1 (the left panel), the distribution of window sizes on the X chromosome regions is given when the distance between any adjacent SNPs in a region cannot exceed 200 kb. The horizontal axis represents the number of SNPs in a region and the vertical axis shows the number of regions of a given size.

To obtain the empirical P-values of the 77 HS association tests, we randomly permuted the phenotypic status and calculated these test statistics accordingly for 50 000 times. The P-value of an HS test statistic is then estimated. This permutation procedure was done only on the regions identified by the single-marker analysis. Thus, it did not take into account the single-marker analysis in the screening of all 1804 SNPs. The results are presented in Figure 2. The horizontal line is

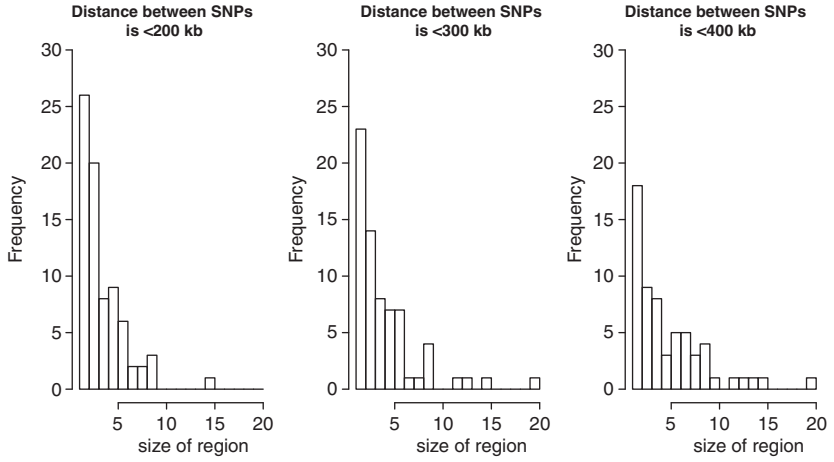

Figure 1 Distribution of the sizes of X chromosome regions.

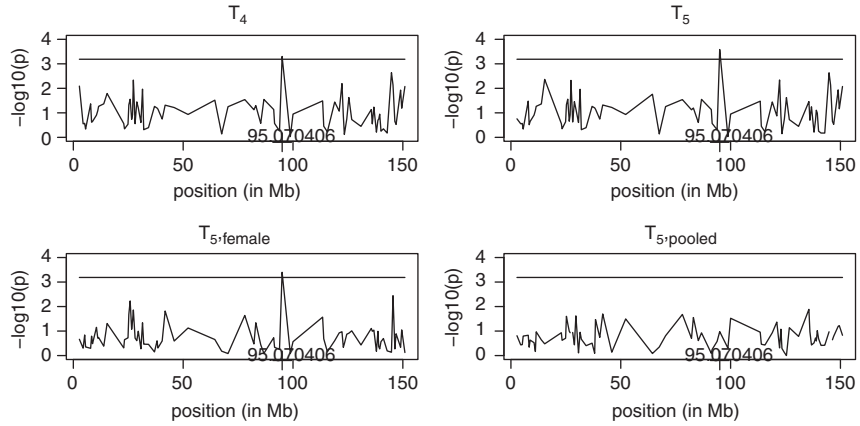

Figure 2 The *P*-values of the HS association tests on the X chromosome.

Table 1 The five SNPs in the disease-preventive haplotype *ATGAC* for AMD

			Freq. of first allele in			*Nominal* P*-values*		
RS number	*Position*	*Alleles*[a]	*Males*	*Females*	*Pooled*	H_1[b]	H_1[c]	*Assoc.*[d]
rs707289	95940577	A/G	0.58	0.74	0.69	0.021	0.515	0.0127
rs10521496	96104694	T/C	0.48	0.60	0.56	0.140	0.037	0.000009
rs10521495	96123338	C/G	0.10	0.14	0.13	0.401	0.649	0.0251
rs1886894	96212007	A/G	0.48	0.47	0.47	0.947	0.124	0.112
rs1012930	96212199	T/C	0.52	0.53	0.52	0.954	0.197	0.122

Abbreviations: AMD, age-related macular degeneration; SNP, single nucleotide polymorphism.
[a]Based on HapMap data.[16]
[b]H_1: Equal allele frequency in males and females.
[c]H_2: HWE holds in females.
[d]Testing association.[1]

the threshold of $\alpha=0.05/77=6.5\times10^{-4}$ by Bonferroni correction. With this α level, there was a chromosomal region associated with AMD. It consists of five SNPs: rs707289, rs10521496, rs10521495, rs1886894, and rs1012930, including SNP rs10521496. The nominal *P*-values of T_4, T_5, and $T_{5,\text{female}}$ (k=4 or 5) on this region are 5×10^{-4}, 2.6×10^{-4}, and 4×10^{-4}, respectively. Their corresponding Bonferroni-corrected *P*-values are 0.0385, 0.0200, and 0.0308. When $T_{5,\text{pooled}}$ was used, there was no region associated with AMD after Bonferroni correction. For each of the five SNPs in this haplotype, the allele frequencies in males, females and pooled samples, *P*-values for testing whether or not males and females have equal allele frequency and whether or not HWE holds in females, and *P*-values for association are given in Table 1.

We also examined the distribution of window sizes (the middle and right panels of Figure 1), when the maximum distance between any adjacent SNPs in a region cannot exceed 300 and 400 kb, and the corresponding HS tests (T_5) (Figure 3). Although we chose 200 kb as the maximum distance between any adjacent SNPs in a window to report our results, Figures 2 and 3 show that results are robust to the choices of distances.

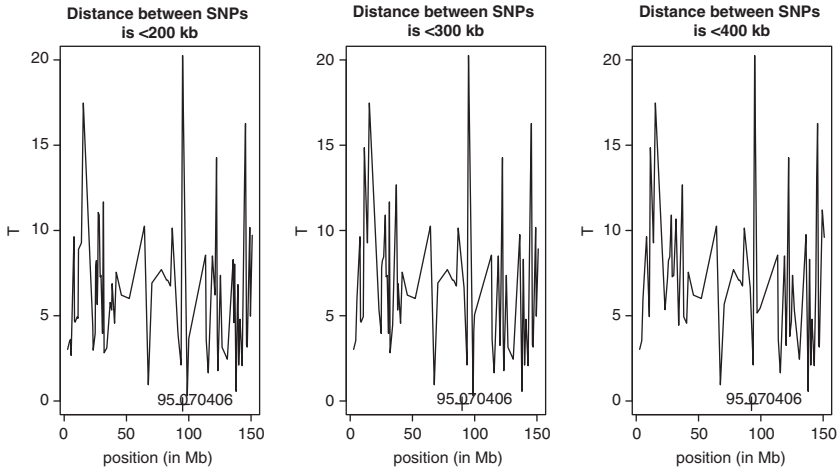

Figure 3 The HS statistic (T_5) on the X chromosome with different maximum distances between SNPs within X chromosome regions.

Table 2 Odds ratios and PAR% for the haplotype *ATGAC* on the five SNPs on the X chromosome

	Female		
	Dominant	*Recessive*	*Male*
Odds ratio (95% CI)	0.2105 (0.063, 0.701)	0.15 (0.04, 0.568)	2.2 (0.687, 7.045)
PAR% (95% CI)	−23.58 (−41.9, −5.3)	−9.69 (−18.3, −1.1)	−21.05 (−50.7, 8.6)

Abbreviations: CI, confidence interval; PAR%, population attributable risk percentages; SNP, single nucleotide polymorphism.
The disease preventive haplotype for females is *ATGAC* with frequency 0.276 in HapMap CEU.

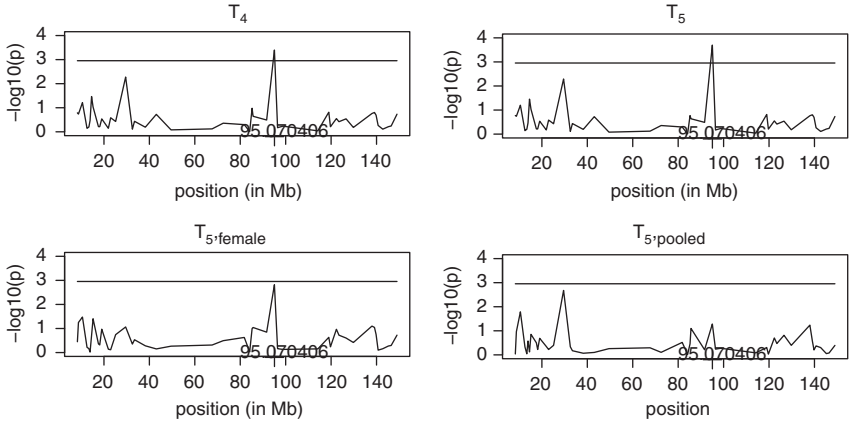

Figure 4 The *P*-values of HS test on the X chromosome using training and testing data sets.

Training-testing sets approach

In the previous analysis, there is no independent dataset for a replication study, although SNP rs10521496 has *P*-value $<0.05/1804$ in the single-marker analysis.[1] Hence, we also applied the training-testing sets approach. When applying this approach, we randomly chose a quarter of cases and half of controls (24 females with 13 cases, and 25 males with 12 cases). The rest of samples formed the testing set. We used the training set to calculate the single-marker statistic (sum of the two one-degree-of-freedom χ^2-tests in females and males; the last one in Table 2 of Zheng *et al*[1]) at each of the 1804 SNPs. We then chose the top 20% of the SNPs as tagging SNPs (360 SNPs) to form regions, with which HS association tests are applied using the testing set. When the distance between any adjacent SNPs in a region does not exceed 200 kb, 86 regions on the X chromosome that contained more than one SNP were obtained. We then carried out HS analysis using the testing set. The four HS tests: T_4, T_5, $T_{5,\text{female}}$, and $T_{5,\text{pooled}}$ were calculated. The empirical *P*-values of these tests were also obtained by 50 000 permutations. On the basis of T_4 and T_5, the most significant X chromosome region contained the same five SNPs: rs707289, rs10521496, rs10521495, rs1886894, and rs1012930 as before. The nominal *P*-values of T_4 and T_5 are 0.00036 and 0.00028, respectively (0.03096 and 0.02408 after Bonferroni correction). The nominal *P*-values of $T_{5,\text{female}}$ and $T_{5,\text{pooled}}$ are 0.0014 and 0.04862, which are not significant after Bonferroni correction. These results are given in Figure 4, where the horizontal line is the threshold of $\alpha=0.05/86=5.8\times10^{-4}$ by the Bonferroni correction. The results from using training-testing sets approach are consistent with those from the analysis using HS association tests based on all samples.

Odds ratios and PAR% for haplotypes

To identify haplotypes associated with AMD, we chose the window with the strongest signal with the five SNPs (rs707289, rs10521496, rs10521495, rs1886894, and rs1012930), and calculated the odds ratios and population attributable risk percentages (PAR%) of the haplotypes in this window. A formal definition of PAR% is given below using a dominant haplotype (H) as an example. Suppose there are a cases and b controls with at least one copy of H (exposed), and c cases and d controls without H (unexposed). Then the incidence in the exposed is $I_e=a/(a+b)$ and the incidence in the unexposed is $I_u=c/(c+d)$ with the incidence in the population $I_p=(a+c)/(a+b+c+d)$. The PAR is defined as I_p-I_u and the PAR% is given by $100(1-I_p/I_u)$, which is the percentage of the incidence of the disease in the total population that would be eliminated if the exposure were eliminated. Using the AMD data, the PAR% measures the proportion of AMD cases in the total population that would be preventable if the haplotype *ATGAC* were absent. A negative value of PAR% indicates that the haplotype is disease preventive (which is equivalent to odds ratio less than one). The results of odds ratios and PAR% are reported in Table 2. Females and males were analyzed separately. The dominant model in females compares individuals having at least one given haplotype with those who have no such haplotype. The recessive model in females compares individuals having two copies of the given haplotypes with those who have at most one such haplotype. For males, it compares individuals with and without this haplotype. A haplotype *ATGAC* is found to be disease preventive among females. The frequency of *ATGAC* in HapMap CEU is 0.276, and it is the most common haplotype over these five SNPs. Under the female dominant model, the odds ratio is 0.2105 with a 95% confidence interval (0.063, 0.701). In other words, among females without this haplotype, the likelihood of AMD is increased by a factor of 4.75 with a 95% confidence interval (1.43, 15.82). The recessive model in females showed a similar pattern. Among females, the χ^2-statistic of comparing the frequency of *ATGAC* with the frequency of all other haplotypes combined is 13.57, and its nominal P-value is 0.00023. These results are given in Table 3. The PAR% among females under the dominant model is −23.58% with a 95% confidence interval (−41.9%, −5.3%), which presents a strong evidence that this haplotype is disease preventive. Under the recessive model among females, the value of PAR% is similar. The role of this haplotype among males is not clear because the odds ratio is 2.2 with a 95% confidence interval (0.687, 7.045). The value of PAR% among males does not provide a clear indication.

We also checked the odds ratio and PAR% for the next most significant window of another five SNPs: rs10521866, rs4824253, rs10521868, rs5904817, and rs10521869 from the analysis when all case–control samples were used. The P-values of the HS statistics (T_4, T_5 and $T_{5,\text{female}}$) of this window are the second lowest, but not significant after Bonferroni correction. The odds ratio shows that the haplotype *GACAT* on the above five SNPs is disease preventive among females. Its frequency in HapMap CEU is 0.578. The effect of this haplotype among males is not clear either. These results are given in Table 4.

Table 3 The frequencies of haplotype *ATGAC* and all other haplotypes combined on the five SNPs on the X chromosome

	Case	*Control*
ATGAC	0.273	0.586
Other haplotypes	0.727	0.414

Abbreviation: SNP, single nucleotide polymorphism.
The nominal P-value of the χ^2-statistic is 0.00023.

Pairwise LD

The pairwise LD of the five SNPs in *ATGAC* was examined using Haploview using HapMap CEU data. These five SNPs do not belong to the same LD block. We searched SNPs on the X chromosome in HapMap and found there were 206 SNPs in HapMap in this region, which were not genotyped in the 100K SNPs. The pairwise LD of the next five SNPs in *GACAT* was also examined. These SNPs do not form any LD blocks either.

DISCUSSION

The X chromosome contains rich information about population history and linkage disequilibrium.[12,13] We proposed simple haplotype-sharing association tests to analyze the X chromosome and applied the results to the subset of NEI AREDS. We studied two strategies for analysis of the X chromosome in GWAS. First, we applied a single-marker analysis to rank all the SNPs on the X chromosome, and the top-ranked SNPs are selected. Then, haplotype analysis is carried out on the regions built on the selected SNPs. The multiple testing for haplotype analysis is controlled at the second stage, rather than the first stage. The second approach is to use the training-testing sets. Randomly draw cases and controls to form a training set, using which we conducted a single-marker analysis to select the top SNPs. Then, we used the remaining case–control samples (the testing set) to conduct haplotype analysis. As training-testing tests are independent, the testing set analysis can be regarded as independent validation of results from using the training set. Likewise, we only control Type I error at the second stage using the testing set. When independent replication samples are available, the first approach should be more powerful as all samples are used in both single-marker and haplotype analyses. In addition, if single-marker analysis is also significant on the same region where haplotype analysis is significant, the type I error seems to be controlled if not conservative. However, if there is no replication study and especially the regions are selected based on the ranks of P-values, the training-testing sets approach is more appropriate. We believe, for haplotypes with true associations, both approaches should result in consistent conclusions.

Table 4 Odds ratios and PAR% for the haplotype *GACAT* on the X chromosome

	Female		
	Dominant	*Recessive*	*Male*
Odds ratio (95% CI)	0.2357 (0.062, 0.892)	0.1625 (0.052, 0.508)	0.4675 (0.175, 1.251)
PAR% (95% CI)	−26.14% (−45.72%, −6.56%)	−13.59% (−24.33%, −2.85%)	15.37% (−4.78%, 35.51%)

Abbreviations: CI, confidence interval; PAR%, population attributable risk percentages.
The disease preventive haplotype for females is *GACAT* with frequency 0.578 in HapMap CEU.

Our haplotype analysis of the X chromosome from the GWAS for AMD confirmed the results of the single-marker analysis on the X chromosome for AMD.[1] Moreover, a chromosomal region consisting of five SNPs (rs707289, rs10521496, rs10521495, rs1886894, and rs1012930) has been identified as associated with AMD. A disease-preventive haplotype (*ATGAC*) on these five SNPs was also identified. This region is covered by the gene *DIAPH2*, which is known to cause premature ovarian failure (POF) in females. Our results indicated that *DIAPH2* may be associated with both premature ovarian failure (POF) and AMD. Although clinical studies have reported the relation between POF and some eye disease symptoms,[14] the relation between POF and AMD has not been reported.[15] Our finding may motivate future clinical and genetic studies for this candidate gene.

As there are 206 SNPs in the haplotype *ATGAC* in HapMap[16] that were not genotyped in the 100K SNPs, fine mapping and analyses of these SNPs using existing or new case–control samples would further help to localize disease loci on the X chromosome. Currently, GWAS uses 500K to 1 million SNPs. Hence the X chromosome would contain about 5–10 times more SNPs than that with 100K SNPs. Applying our approach to current GWAS would help to identify much more candidate-regions or genes for further investigation than single marker analysis. Although we focused on a haplotype-sharing approach for the X chromosome, comparing this approach to other methods for haplotype analysis for the X chromosome under various genetic models requires future work.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

The use of NEI AREDS GWAS data was approved by the Data Access Committee of NEI through dbGap. We would like to thank a reviewer for the helpful comments and suggestions, especially on the two-stage analysis. We would also like to thank Dr James Troendle on discussion of multiple testing issues.

1 Zheng G, Joo J, Zhang C, Geller NL: Testing association for markers on the X chromosome. *Genet Epidemiol* 2007; **31**: 834–843.
2 Clayton DG: Testing for association on the X chromosome. *Biostatistics* 2008; **9**: 593–600.
3 Tzeng JY, 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* 2003; **72**: 891–902.
4 Zhang SL, Sha Q, Chen H, Dong J, Jiang R: Transmission/disequilibrium test based on haplotype sharing for tightly linked markers. *Am J Hum Genet* 2003; **73**: 566–579.
5 Excoffier L, Slatkin M: Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995; **12**: 921–927.
6 Li Y, Sung WK, 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.
7 Clayton DG, Jones H: Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet* 1999; **65**: 1161–1169.
8 Bourgain C, Genin E, Quesneville H, Clerget-Darpoux F: Search for multifactorial disease susceptibility genes in founder populations. *Ann Hum Genet* 2000; **64**: 255–265.
9 Toivonen HT, Onkamo P, Vasko K *et al*: Data mining applied to linkage disequilibrium mapping. *Am J Hum Genet* 2000; **67**: 133–145.
10 Frieman DS, O'Colmain BJ, Munoz B *et al*: Prevalence of age-related macular degeneration in the United States. *Arch Ophthalmol* 2004; **122**: 564–572.
11 Klein RJ, Zeiss C, Chew EY *et al*: Complement factor H polymorphism in age-related macular degeneration. *Science* 2005; **308**: 385–389.
12 Schaffner SF: The X chromosome in population genetics. *Nat Rev Genet* 2004; **5**: 43–51.
13 Laan M, Wiebe V, Khusnutdinove E, Remm M, Paabo S: X chromosome as a marker for population history: linkage disequilibrium and haplotype study in Eurasian populations. *Eur J Hum Genet* 2005; **13**: 452–462.
14 Smith JA, Vitale S, Reed GF *et al*: Dry eye signs and symptoms in women with premature ovarian failure. *Arch Ophthalmol* 2004; **122**: 151–156.
15 de Jong PTVM: Age-related macular degeneration. *N Engl J Med* 2007; **355**: 1474–1485.
16 Thorisson GA, Smith AV, Krishnan L, Stein LD: The International HapMap Project Web site. *Genome Res* 2005; **15**: 1591–1593.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)