

Comparison of SNP Tagging Methods Using Empirical Data: Association Study of 713 SNPs on Chromosome 12q14.3–12q24.21 for Asthma and Total Serum IgE in an African Caribbean Population

Peter B. Chi,^{1,2} Priya Duggal,³ W.H. Linda Kao,¹ Rasika A. Mathias,³ Audrey V. Grant,^{1,2} Maria L. Stockton,² Joe G.N. Garcia,⁴ Roxann G. Ingersoll,⁵ Alan F. Scott,⁵ Terri H. Beaty,¹ Kathleen C. Barnes,^{2*} and M. Daniele Fallin^{1,6}

¹Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland

²Division of Allergy and Clinical Immunology, Department of Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland

³Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland

⁴Division of Pulmonary and Critical Care Medicine, Department of Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland

⁵McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland

⁶Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland

Few comparison studies have been performed on single nucleotide polymorphism (SNP) tagging methods to examine their consistency and effectiveness in terms of inferences about association with disease. We applied several SNP tagging methods to SNPs on chromosome 12q ($n = 713$) and compared the utility of these methods to detect association for asthma and serum IgE levels among a sample of African Caribbean families from Barbados selected through asthmatic probands. We found that a high level of information regarding association is retained in Clayton's htSNP, Stram's TagSNP, and de Bakker's Tagger. We also found a high degree of consistency between TagSNP and Tagger. Using this set of 713 SNPs on chromosome 12q, our study provides insight towards analytic strategies for future studies of complex traits. *Genet. Epidemiol.* 2006. © 2006 Wiley-Liss, Inc.

Key words: LD blocks; SNP tags; FBAT; association

Contract grant sponsor: National Institutes of Health Grants; Contract grant numbers: 001H266615 and A150024-03; Contract grant sponsor: Asthma of Allergy Foundation of America New Investigator Award Grant; Contract grant sponsor: Mary Beryl Patch Turnbull Scholar Program; Contract grant sponsor: Intramural Research Program of the National Human Genome Research Institute, NIH.

*Correspondence to: Kathleen Barnes, Johns Hopkins Asthma & Allergy Center—3A.62A, 5501 Hopkins Bayview Circle, Baltimore, MD 21224. E-mail: kbarnes@jhmi.edu

Received 15 March 2006; Revised 25 April 2006; Accepted 22 May 2006

Published online in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/gepi.20172

INTRODUCTION

Questions of cost versus effectiveness are of great importance in genetic epidemiology. Studies generally require a large amount of genotyping, which has become more efficient in recent years but remains an expensive process. Furthermore, current human genome databases, such as the UCSC Genome Browser, are continually growing, making additional markers available each year, and moving toward genotyping projects on a scale of 500,000 or more markers per participant [Karolchik et al., 2003]. As new methods are developed to accommodate such large numbers, analytical questions become more complex.

The simplest genetic markers are single nucleotide polymorphisms (SNPs), in which one DNA

base is substituted for another. SNPs are typically utilized in an association study in two ways: as direct and indirect genetic measures. The direct approach considers candidate SNPs, which are suspected of being causal for the disease of interest. Therefore, the genetic feature of interest is directly measured. The indirect approach treats SNPs as markers and exploits linkage disequilibrium (LD) between these markers and (unmeasured) causal SNPs [Clayton et al., 2004]. Thus, entire regions can be genotyped using a set of SNPs without regard to functionality, and a dense map of information is created in anticipation that some of these SNPs will be correlated with a causal mutation, thus showing association with the disease trait of interest. The high cost of genotyping becomes an issue here, since the

indirect approach typically requires many more SNPs than the direct approach [Chapman et al., 2003]. However, knowledge of a biological mechanism for any one SNP is not trivial, and is unknown for the vast majority of polymorphisms. For this reason, indirect approaches are often much more practical than direct approaches.

To deal with the issue of genotyping costs, SNP tagging methods have been developed. In regions of high LD, where many SNPs are frequently inherited together and thus highly correlated within populations, it may not be necessary to genotype all of the SNPs in a given region to capture all of the genetic information. In fact, it would be wasteful to do so. A reduced set of SNPs can be representative of most, if not all, genetic variation in that region. The goal of SNP tagging methods is to determine which SNPs to include in this reduced set [Stram et al., 2002].

To date, several tagging methods have been proposed. The basis for each method can roughly fall into one of two different categories: structured and unstructured [Forton et al., 2005]. In the structured approach, regions of high LD are defined into "blocks" through various algorithms [Daly et al., 2001; Johnson et al., 2001; Patil et al., 2001; Gabriel et al., 2002; Zhang and Jin, 2003; Barrett et al., 2005], and tagging SNPs are then identified to represent each block. Conversely, the unstructured approach does not define LD blocks, but rather aims to identify a subset of markers that represent the entire region in question [Stram et al., 2002; Chapman et al., 2003; Ke and Cardon, 2003; Sebastiani et al., 2003; de Bakker et al., 2005]. While comparisons between methods have been made in terms of agreement between SNPs selected [Butler et al., 2005; Duggal et al., 2005; Nicodemus et al., 2005], few comparative studies have examined consistency among these methods in terms of inference about association with disease [Halldórsson et al., 2004]. Because most studies are limited to one SNP tagging method, it is important to consider how these methods perform against each other to detect association in an empirical setting.

Here we examine several commonly employed SNP tagging methods in a set of 713 SNPs from chromosome 12q, genotyped in an African Caribbean, family-based population from Barbados, for two asthma-related phenotypes—asthma susceptibility as a qualitative trait, and the associated quantitative trait, total serum IgE, which has been shown to be linearly related to the log of the odds for self-reported asthma

[Burrows et al., 1989]. Previously, novel evidence for linkage to asthma and atopy was discovered on a large region of 12q using micro-satellite markers [Barnes et al., 1996], and this subsequent high-throughput genotyping project was performed using 12q SNPs in an IlluminaTM platform fine mapping panel (Illumina, San Diego, CA), which was determined to give the best coverage of this region of linkage. We first show association results for all SNPs genotyped in the study, then compare this to the results using only the selected subset of tagging SNPs identified by five different SNP tagging algorithms. Our goal is to show that SNP tagging methods can reduce genotyping effort while retaining association signals for asthma and IgE, but there is variation depending upon the tagging method employed.

MATERIALS AND METHODS

STUDY POPULATION

Barbados is the eastern-most island of the Caribbean, and its population is predominantly of African descent with approximately 25% Caucasian admixture, nearly identical to that observed in African Americans [Nickel et al., 1999]. Due to its high asthma prevalence (estimated at 13% of the general population based on Emergency Department visits and possibly as high as 20%), Barbados is an ideal population for studies of asthma. In this study, African Caribbean subjects were recruited as part of an ongoing asthma genetics study described previously [Barnes et al., 1996, 2006]. A total of 1,088 participants in 125 nuclear and extended families were recruited through probands identified at sites across the island, including six polyclinics, two private clinics, and the Accident and Emergency Department (A&ED) of Queen Elizabeth Hospital (QEH). All participants gave their informed consent under a study protocol approved by the Johns Hopkins Medical Institutions' Joint Committee on Clinical Investigations (JCCI) and the Ministry of Health in Barbados. Of the 1,088 recruited individuals, 668 were genotyped for this study based on DNA quality, sample availability, and other quality control issues such as sample labeling discrepancies. These 668 individuals were not significantly different than those not included with regard to asthma status, age, or any other relevant variables.

Asthma was defined as previously described [Barnes et al., 1996; Zambelli-Weiner et al., 2005],

as both: (1) a reported history of asthma, (2) a history of physician-diagnosed asthma (past/current), and (3) a history of wheezing without an upper respiratory infection (URI) or two out of the following four symptoms: wheezing with a URI, cough without a URI, shortness of breath, and tightness in the chest. Total IgE concentrations were measured as described previously [Barnes et al., 1996]. IgE values were log-transformed to minimize skewness and then adjusted for age and sex.

GENOTYPING

DNA was extracted from peripheral blood samples using standard protocols. Whole genome amplification (WGA) of DNA was prepared by OmniPlex Technology (Rubicon Genomics Inc., Ann Arbor, MI). A total of, 1,261 SNPs on chromosome 12q14.3–12q24.21 (dbSNP build 123) were attempted on the IlluminaTM BeadArrayTM platform (Illumina). These included 735 markers from fine-mapping panel 42 (constructed with the goal of providing coverage for linkage based on SNPs validated in a European population), supplemented with custom SNPs chosen in 2003, prior to HapMap data availability. Assays were completed for 1,083 SNPs, but only 713 had $MAF > 0.01$ and met Hardy-Weinberg proportions in founders. This loss of SNPs is likely due to two problems: first, that there was little information about genomic polymorphism available when these SNPs were chosen, and second, that SNP selection was based on samples of European descent. Current strategies would use more recent SNP data focused on African samples.

DATA SEGMENTATION

To create manageable sets of SNPs for analysis, SNPs were first parsed into three regions separated by two large gaps in the panel (each over 12 Mb), resulting in just over 200 SNPs in each region. However, because two of the SNP tagging programs rely on haplotype frequencies (TagSNP and htSNP), which become less meaningful over such large distances that show no LD, these three regions were further broken down into “analytic segments” of less than 20 SNPs each. To accomplish this, regions were first split at every gap ≥ 200 kb between consecutive SNPs. Then, for segments that still contained > 20 SNPs, an additional split was made at spots of low pairwise-LD ($D' \leq 0.70$, but generally much closer to 0). This resulted in 76 analytic segments whose

size and LD characteristics are shown in Figure 1. Segment sizes ranged from 3 to 19 SNPs (panel a), and the length of the segments ranged from roughly 100 to 1,200 kb (panel b). SNP density ranged from 1 SNP per 10 kb to 1 per 100 kb (panel c). As expected, segments with higher SNP density had higher median LD, since SNPs in close proximity are more likely to be in strong LD (panel d).

METHODS COMPARED

Five SNP tagging methods were considered, but after accounting for differing options for each method, a total of eight methods were implemented on these asthma data. First, single SNP association tests across the entire set of 12q SNPs were performed using FBAT (method 1, see below). Second, each SNP tagging algorithm was used to define SNP subsets for analysis: structured SNP tagging with blocking according to Gabriel et al., (method 2); structured SNP tagging with blocking according to solid spine of LD (SSLD) (method 3); and unstructured SNP tagging using Stram's TagSNP (methods 4 and 5); unstructured SNP tagging using Clayton's htSNP (methods 6 and 7); and unstructured SNP tagging using de Bakker's Tagger (method 8).

FBAT GENETIC ASSOCIATION ANALYSIS

The Family-Based Association Test (FBAT) as implemented in the FBAT software was used to test association between each SNP and asthma or $\text{Log}[\text{IgE}]$ [Rabinowitz and Laird, 2000; Laird et al., 2000; Horvath et al., 2001], assuming an additive model and hypothesis testing for linkage or association.

SNP TAGGING METHODS

Gabriel blocking. In this structured tag SNP selection method, available in Haploview version 3.11 [Barrett et al., 2005], LD blocks are first defined according to the method of Gabriel et al. [2002] and SNPs that represent the underlying haplotypes are chosen within blocks. Pairs of SNPs are defined as having “strong evidence for historical recombination” if the upper confidence bound on D' is less than 0.9, and a block is defined as a region over which less than 5% of SNP pairs show strong evidence of historical recombination. Tagging SNPs are then selected per block as a set of SNPs which define all haplotypes above a given frequency threshold (1% in this report) (method 2).

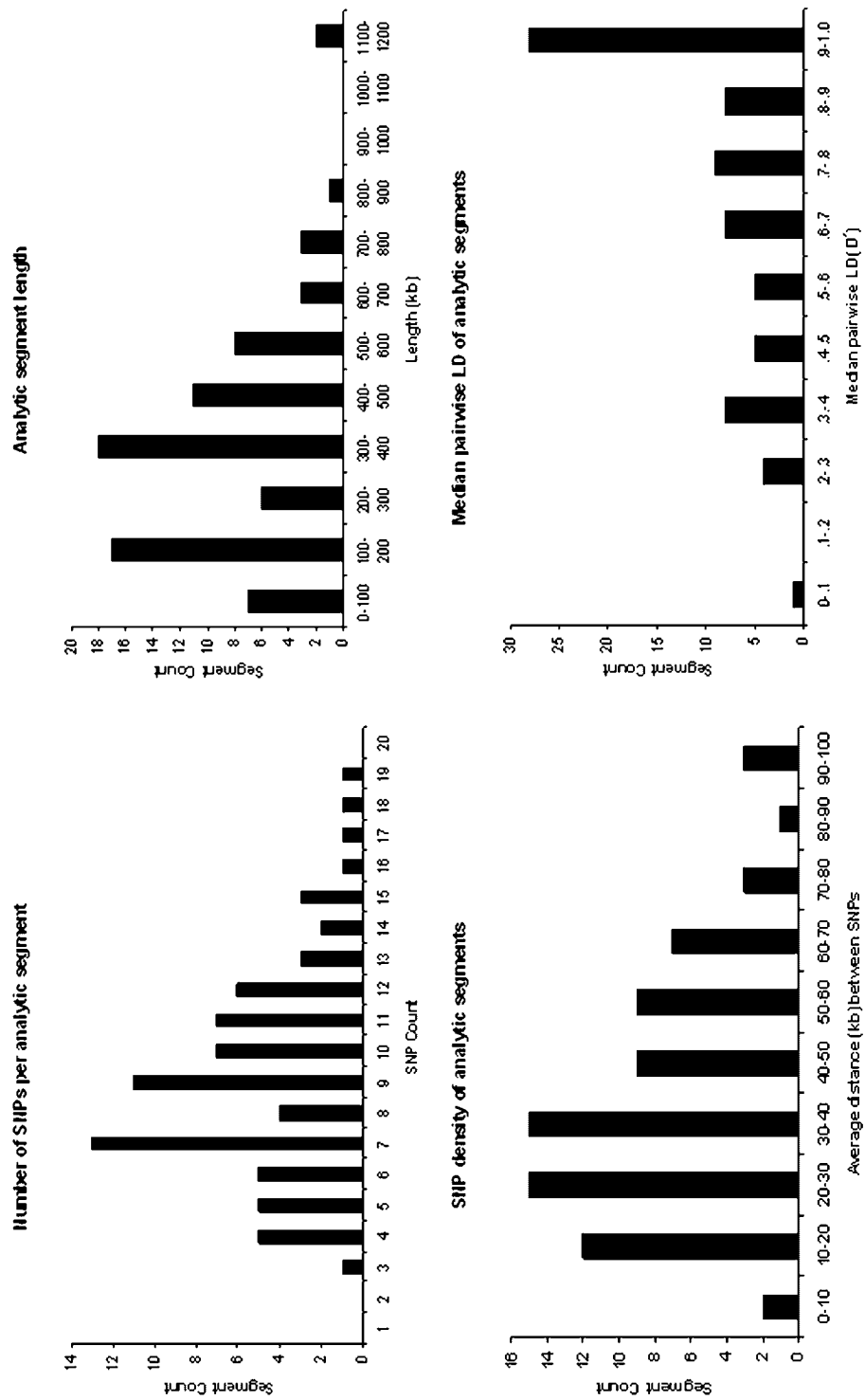


Fig. 1. Descriptive characteristics of analytic segments in the dataset.

SSLD blocking. This structured tagging method is also available in HaploView and uses the SSLD algorithm to define blocks by searching for a “spine” of LD, such that the first and last markers in a block are in strong LD with all intermediate markers. However, the intermediate markers are not necessarily in LD with each other. Tagging SNPs are then selected per block to define all haplotypes $>1\%$ frequency (method 3).

htSNP. This unstructured approach selects SNPs from an entire region of interest, without regard to LD blocks [Clayton, 2001]. Tagging SNPs are selected based on the ability of any SNP subset to maintain the overall haplotype diversity observed when considering all SNPs. We relied on the measure r^2 , at a threshold of 0.8, to determine maintenance of diversity. This is a measure of the proportion of information lost by using each possible set of SNPs compared to the full set. The htSNP method requires the input of haplotype frequency estimates, which were calculated through the SNP HAP software, version 1.3.1 [Clayton, 2005]. In an attempt to maintain consistency for the purpose of a comparison with Stram’s TagSNP (below), two approaches were employed based on cumulative estimated haplotype frequencies: htSNP(60) attempts to capture 60% of the cumulative frequencies (method 4), and htSNP(100) attempts to capture 100% of the haplotypes (method 5).

TagSNP. TagSNP also uses an unstructured approach to identify the SNP subset that best captures the complete SNP information. Because problems may arise if there are multiple rare haplotypes, the approach written for the TagSNP program attempts to avoid this by optimizing the predictability of only the common haplotypes. For common haplotypes, the “haplotype dosage,” which is the count of the number of copies of a haplotype for each individual, is estimated using all SNPs. This is then repeated using only subsets of SNPs, and the squared correlation between the true and predicted haplotype dosage (R_h^2) for all individuals is calculated to find the subset of SNPs with the highest correlation to the full information. Two TagSNP implementations are used, again based on cumulative haplotype frequencies for comparison with htSNP (above): TagSNP(60) captures 60% of cumulative frequencies (method 6), and TagSNP(100) captures 100% of the haplotypes (method 7).

Tagger. This is also an unstructured method, where bins of SNPs are designed across a region according to pairwise LD. Bins of SNPs are

created based on the specified r^2 threshold, then one SNP is selected to represent the remainder of SNPs in that bin [de Bakker et al., 2005]. In this study, we used an r^2 threshold of 0.8 and minimum allele frequency of 0.05 (method 8).

RESULTS

A total of 51 SNPs from the full dataset were associated with asthma and 45 SNPs were associated with Log[IgE], at the nominal $P < 0.05$ level (see top panels of Fig. 3). The strongest signal for asthma was observed near the gene MYF6, with $P = 0.002$ and a corresponding $P = 0.02$ for Log[IgE]. The strongest signal for Log[IgE] was observed near MGAT4C, with $P = 0.001$ and a corresponding $P = 0.004$ for asthma. There was an overlap of 23 SNPs which showed $P < 0.05$ for both phenotypes.

SNPs chosen by each of these SNP tagging methods are shown for three example segments in Figure 2, demonstrating the pattern of selection for segments representing high LD, medium LD, and low LD. With these tagging SNPs, single-SNP results were compared with SNPs selected by each method to results from all genotyped SNPs. The goal was to identify a method that drops the greatest number of SNPs while losing the least information about association to the traits of interest.

There was a modest degree of overlap in SNP selection between different methods, especially in regions of high LD (Fig. 2). As expected, all methods retained more SNPs as LD decreases. In general, the unstructured methods retain fewer markers than the structured approaches, presumably because they are not restricted to full representation of each LD block, but rather combine information across the entire segment. Figure 3 illustrates the FBAT association results for individual SNPs selected by each method compared to the complete set. The proportion of independent signals in the complete dataset captured by each SNP tagging subset is shown in Table I. Because some SNPs are in high LD with each other, the proportions of total associated SNPs retained by each tagging method may not be as important as the proportion of associated signals retained under each method. To address this, we define each contiguous set of significant SNPs with high pairwise LD ($D' > 0.8$) as one independent signal; this yields 42 independent signals of 51 for asthma, and 39 independent

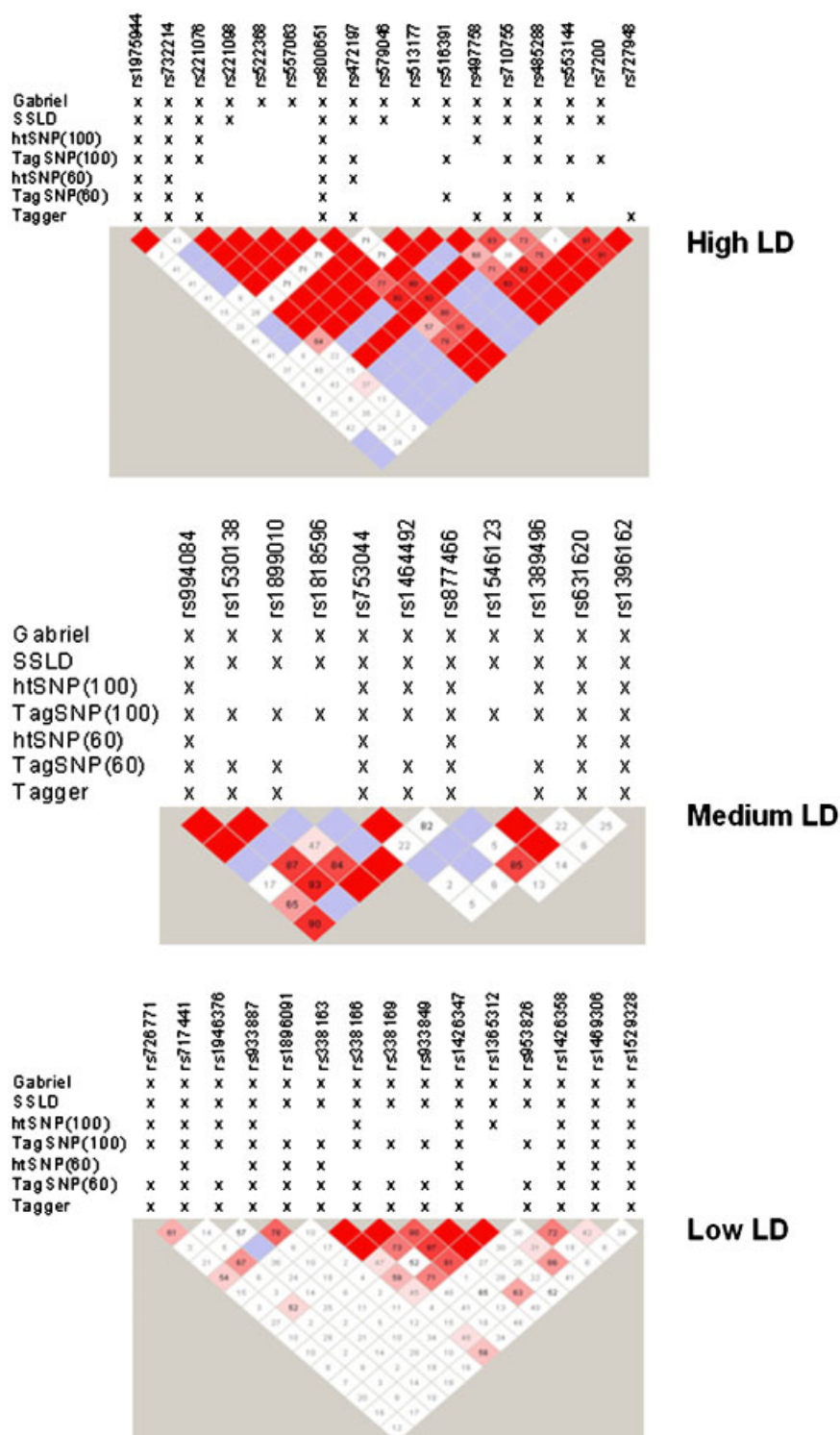


Fig. 2. LD plots and tagging SNPs for each method from three selected segments. Illustration of SNPs selected by each method on three representative analytic segments, for High LD (avg. $D' = 0.86$), Medium LD (avg. $D' = 0.75$), and Low LD (avg. $D' = 0.59$). SNPs selected by each method are indicated with an 'x.'

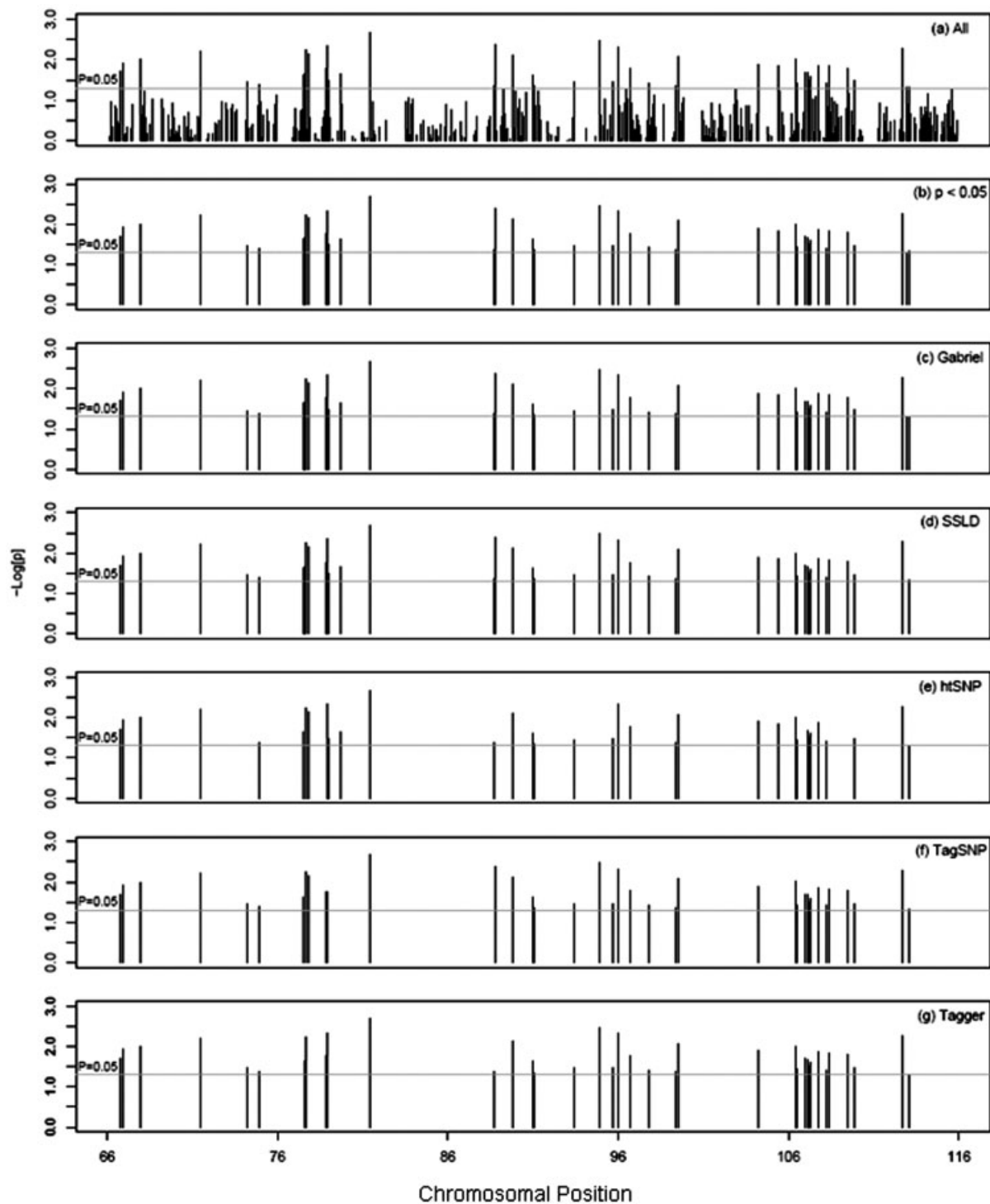


Fig. 3. Graphical representation of the $-\log(P\text{-value})$ for FBAT tests of association with asthma. (a) all 713 SNPs, (b) all 51 SNPs with $P\text{-value} < 0.05$, (c-g) SNPs with $P\text{-value} < 0.05$ that were retained in the subset of tagging SNPs selected by each method as indicated.

signals of 45 for Log[IgE]. The proportion of these signals captured (defined as the retention of at least one significant SNP per independent signal) by each SNP tagging subset is shown in bold in

Table I. These proportions were not sensitive to our designation of LD threshold.

The differences across methods are substantial throughout the full dataset. Both the Gabriel and

TABLE I. Haplotype-tagging SNPs for all methods and associations with Asthma and Log[IgE]

	Complete set	Gabriel	SSLD	htSNP(100)	TagSNP(100)	htSNP(60)	TagSNP(60)	Tagger
<i>Asthma</i>								
SNPs in selected subset	713	696	681	572	677	369	590	595
(Proportion of complete set)		(0.98)	(0.96)	(0.80)	(0.95)	(0.52)	(0.83)	(0.83)
SNPs $P < 0.05$	51	50	49	38	49	25	42	42
(Proportion of complete set)		(0.98)	(0.96)	(0.75)	(0.96)	(0.49)	(0.82)	(0.82)
Independent SNPs $P < 0.05$	42	42	42	37	42	21	38	37
(Proportion of complete set)		(1.00)	(1.00)	(0.88)	(1.00)	(0.50)	(0.90)	(0.88)
<i>Log[IgE]</i>								
SNPs in selected subset	713	696	681	572	677	369	590	595
(Proportion of complete set)		(0.98)	(0.96)	(0.80)	(0.95)	(0.52)	(0.83)	(0.83)
SNPs $P < 0.05$	45	43	40	35	41	17	36	36
(Proportion of complete set)		(0.96)	(0.89)	(0.78)	(0.91)	(0.38)	(0.80)	(0.80)
Independent SNPs $P < 0.05$	39	39	38	34	36	14	35	35
(Proportion of complete set)		(1.00)	(0.97)	(0.87)	(0.92)	(0.36)	(0.90)	(0.90)

Note: Our dataset contains a total of 713 SNPs, of which 51 are significantly associated with Asthma, and 45 are nominally associated with Log[IgE] ($P < 0.05$). SNPs were selected by each program from the full set of 713. We then examined the subset of SNPs with $P < 0.05$ for each trait, and the number of SNPs selected by each program from this subset. Our ultimate interest is in the adjusted proportion, which is based on the total number of “signals” detected. Each signal is defined as a contiguous set of SNPs with high LD ($D' > 0.80$), and a method was determined to have captured each signal if at least one SNP from the signal was retained as a tag SNP.

SSLD methods and TagSNP(100) maintain a very large proportion of all SNPs that demonstrated association in the data. However, they do not drop many SNPs overall (retaining 98%, 96%, and 95%, respectively). Conversely, htSNP(60) dropped the greatest number of total SNPs (344 of the total 713), while also dropping a large proportion of SNPs found to be associated with the two traits of interest (capturing only 50% of signals for asthma, and 36% for Log[IgE]). A balance was achieved by htSNP(100), TagSNP(60), and Tagger, all dropping a more modest amount of SNPs (retaining 80%, 83%, and 83%, respectively) while retaining a higher proportion of signals for asthma (88%, 90%, and 88%).

DISCUSSION

The aim of this study was to compare different SNP tagging methods on a real dataset to identify potential inconsistencies in inferences drawn from an association study depending on the method of selecting tagging SNPs. Ideally, one would like to reduce genotyping efforts to an acceptable subset of SNPs, but not if this results in critical loss of information regarding association with a trait of interest (i.e., asthma). For this reason, we focused on agreement in ability to detect association rather than overlap in particular SNPs chosen by each method.

In these applied data, the Gabriel and SSLD structured methods and the TagSNP(100) unstructured method preserve a very large proportion of the asthma-associated SNPs in our data. However, because they do not drop many SNPs overall, the financial benefit from these methods would be modest. Conversely, because htSNP(60) drops the greatest number of total SNPs, this would save the greatest in genotyping costs; however, a large proportion of the asthma-associated SNPs were also dropped. Considering this, the three methods that achieved the best balance were htSNP(100), TagSNP(60), and Tagger. Each captures a reasonable proportion of signals (~90% for both asthma and Log[IgE]), while dropping many SNPs overall. These are potentially the most useful methods for single-SNP analyses. Tagger is becoming the most commonly used, given its implementation in the HapMap project. It is computationally the simplest and does not require prior segmentation of the data.

There are several methodological issues that arose during this process. First is the need to create analytic segments. An even parsing of SNPs per segment, while logistically simple and computationally easy, would not have been ideal in this analysis for several reasons. First, this would cause breaks to occur in some of the areas of high LD, which would undermine the tagging process. Also, as SNP density is too variable across our dataset, an even parsing of SNPs would give rise to short segments where SNP density was high,

and long segments where SNP density was low. For this reason, we chose to divide our segments first based on distance; a division was made at every place where the distance between two consecutive SNPs was ≥ 200 kbp. Then, in segments still containing more than 20 SNPs, an additional break was made at a point of low LD (using a threshold of $D' \leq 0.70$, but was generally much closer to 0). Among our 76 segments, there were only six breaks between markers with pairwise $D' > 0.50$ (five correspond to $r^2 < .03$, one with $r^2 = .32$). These few cross-segment situations of non-zero LD are not likely to affect our results. Also, in light of this issue, it is important to note again that Tagger would not require this prior segmentation.

The comparability across SNP tagging methods is also difficult, given the various options under each method. Because htSNP automatically uses 100% of the estimated haplotype frequencies, it seems logical to use 100% in TagSNP for comparison. However, in segments with many rare haplotypes, TagSNP did not perform as reasonably expected. In this dataset, 40 of the 76 total segments had greater than 20 haplotypes with estimated frequencies of $\leq 1\%$; in these cases, TagSNP tends to retain all SNPs. Thus, from our studies, it appears that it is necessary to limit the haplotypes used by TagSNP. The options are to set a cumulative haplotype frequency threshold, or set a minimum frequency threshold for inclusion of haplotypes to be considered in these algorithms. Given our data, the better choice appeared to be a cumulative haplotype frequency, due to the great variability of haplotype frequency estimates across analytic segments. For example, some segments have only one haplotype with an estimated frequency above 5%. Through a close examination of the data, 60% was chosen as the minimum cumulative haplotype frequency for our analysis, to maximize consistency across segments. Therefore, both TagSNP and htSNP were run either 100% or 60% cumulative haplotype frequency thresholds (for htSNP, this step had to be performed outside of the software). While our goal was to achieve direct comparability, we recognize that what may be appropriate for one program may not be appropriate for another program, and our attempt to maintain consistency may have been unrealistic and unnecessary. As shown in these results, htSNP(100) and TagSNP(60) (our original implementations for each method) yielded the most useful results, with a high proportion of associated SNPs being

retained while dropping a relatively large number of total SNPs.

One general concern for SNP tagging in the context of candidate gene analysis is the focus on indirect association, rather than direct causality; if several SNPs are in high LD, a causal SNP may be dropped in favor of a nearby SNP. Unfortunately there is no way to avoid this; by design, the purpose of tagging methods is to reduce redundancy and capture a minimal set of SNPs to represent all others. The effect could be reduced by forcing inclusion of non-synonymous coding SNPs into the subset. However, if the original SNP sample includes a large percentage of such SNPs, then this process would obviously not be appropriate. Therefore, it is necessary to examine additional SNPs whenever a tag SNP is found to be associated with the trait of interest.

This report includes both haplotype-based and pairwise LD methods for SNP tagging. We chose to compare these with respect to single-SNP association results. One may argue that haplotype analyses would have been more appropriate for SNPs selected from a haplotype-based tagging algorithm, and our results may be slightly different if haplotype analyses were used for comparison of associations. However, this is difficult given the number of different haplotype analysis methods in practice, and the choice of which SNPs to include per haplotype. Ultimately, this may lead to confusion rather than clarity when comparing across tagging methods, so we focused on SNP-based association tests as a standardized comparison.

The greatest limitation in this report is our inability to distinguish truth from observations. We do not know whether the signals detected in this sample reflect true positives nor whether any additional true positives were undetected among our full set of SNPs. Therefore, we are restricted to reporting a comparison of observed findings across tagging methods, without comparing false-positive rates or power. This problem is exacerbated by the use of a fine-mapping linkage panel which was not highly informative for association studies because SNP coverage was not gene-centered, nor was it based on optimal LD coverage. The panel of SNPs only achieved average spacing of 1 SNP per 65 kb, with large variation in spacing despite the attempt to create a uniform distribution in the original construction of the panel. Additional work using better gene coverage must be done to ultimately identify the variants associated with asthma in this sample.

However, these data do provide an empirical set in which to compare tagging methods and strategies for analysis. Further genotyping aimed specifically at candidate genes on 12q is in progress in our sample, with SNPs selected by the Tagger method, based on the work presented here.

Additional concerns include the use of SNPs optimized for Caucasian polymorphisms. One may expect our tagging comparison to change if SNPs chosen for samples of African ancestry were genotyped. For example, these results may be quite different if our panel had been designed based on African SNP information such as the Yoruban data now available through HapMap. However, this was not available at the time of our study, and even now, most of the current SNP panels in use are based on Caucasian polymorphisms. Therefore, our results are relevant until much more work in African-specific populations has been carried out.

This project provided the opportunity to compare these methods on empirical data. Thus, we are able to become fully aware of the differences and consequences that can arise when attempting to use these methods in reality. The number of 12q SNPs in this dataset is large, thus allowing us to perform a thorough investigation of these different methods. Also, all of our SNPs are in the public domain so they may appear in other studies as well, which increases the utility of this report. Our data are part of a large, ongoing genotyping study in Barbados. These conclusions will facilitate future decisions regarding genotyping in this same population, and have provided insight towards other studies involving 12q and asthma, as well as the choice of SNP tagging methods when analyzing real data.

ACKNOWLEDGMENTS

We thank the families and volunteers who graciously agreed to participate in this study. We are especially grateful to Pissamai Maul, Trevor Maul, Kathryn Held, April Zambelli-Weiner, Caitlin Dorer, and Kimberly Donnelly for fieldwork, to Harold Watson, Paul Levett, and Raana Naidu for facilitating the Barbados study, and to William Shao and Shu Zhang for technical assistance. Daniela Baltadjieva and Tiina Berg assisted in preparation of samples and Mark Hansen facilitated genotyping at IlluminaTM. This work was supported by National Institutes of

Health grants U01 HL66615 and AI50024-03, and an Asthma & Allergy Foundation of America New Investigator Award Grant (K.C.B.). K.C.B. was supported in part by the Mary Beryl Patch Turnbull Scholar Program and R.A.M. was supported by the Intramural Research Program of the National Human Genome Research Institute, NIH.

REFERENCES

- Barnes KC, Neely JD, Duffy DL, Freidhoff LR, Breazeale DR, Schou C, Naidu RP, Levett PN, Renault B, Kucherlapati R, Iozzino S, Ehrlich E, Beaty TH, Marsh DG. 1996. Linkage of asthma and total serum IgE concentration to markers on chromosome 12q: evidence from Afro-Caribbean and Caucasian populations. *Genomics* 37:41–50.
- Barnes KC, Freidhoff MS, Nickel R, Chiu YF, Juo SH, Hizawa N, Naidu RP, Ehrlich E, Duffy DL, Schou C, Levett PN, Marsh DG, Beaty TH. 1999. Dense mapping of chromosome 12q13.12–q23.3 and linkage to asthma and atopy. *J Allergy Clin Immunol* 104:485–491.
- Barnes KC, Grant AV, Baltadzhieva D, Zhang S, Berg T, Shao L, Zambelli-Weiner A, Anderson W, Nelson A, Pillai S, Yarnall DP, Dienger K, Ingersoll RG, Scott AF, Fallin MD, Mathias RA, Beaty TH, Garcia JG, Wills-Karp M. 2006. Variants in the gene encoding C3 are associated with asthma and related phenotypes among African Caribbean families. *Genes Immun* 7: 27–35.
- Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21: 263–265.
- Burrows B, Martinez FD, Halonen M, Barbee RA, Cline MG. 1989. Association of asthma with serum IgE levels and skin-test reactivity to allergens. *N Engl J Med* 320:271–277.
- Butler JM, Bishop DT, Barrett JH. 2005. Strategies for selecting subsets of single-nucleotide polymorphisms to genotype in association studies. *BMC Genet* 6(Suppl 1):S72.
- Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. 2004. Mapping complex disease loci in whole-genome association studies. *Nature* 429:446–452.
- Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 74:106–120.
- Chapman JM, Cooper JD, Todd JA, Clayton DG. 2003. Detecting disease associations due to linkage disequilibrium using haplotype tags: A class of tests and the determinants of statistical power. *Hum Hered* 56:18–31.
- Clayton D. 2001. Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci. www.nature.com/ng/journal/v29/n2/extref/ng1001-233-S10.pdf.
- Clayton D. 2005. SNPAP: a program for estimating frequencies of large haplotypes of SNPs (Version 1.3) [<http://www-gene.cimr.cam.ac.uk/clayton/software/>].
- Clayton D, Chapman J, Cooper J. 2004. Use of unphased multilocus genotype data in indirect association studies. *Genet Epidemiol* 27:415–428.
- Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. 2001. High-resolution haplotype structure in the human genome. *Nat Genet* 29:229–232.

- de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. *Nat Genet* 37:1217–1223.
- Deng HW. 2001. Population admixture may appear to mask, change, or reverse genetic effects of genes underlying complex traits. *Genetics* 159:1319–1323.
- Duggal P, Gillanders EM, Mathias RA, Ibay GP, Klein AP, Baffoe-Bonnie AB, Ou L, Dusenberry IP, Tsai YY, Chines PS, Doan BQ, Bailey-Wilson JE. 2005. Identification of tag single-nucleotide polymorphisms in regions with varying linkage disequilibrium. *BMC Genet* 6(Suppl 1):S73.
- Forton J, Kwiatkowski D, Rockett K, Luoni G, Kimber M, Hull J. 2005. Accuracy of haplotype reconstruction from haplotype-tagging single-nucleotide polymorphisms. *Am J Hum Genet* 76:438–448.
- Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The structure of haplotype blocks in the human genome. *Science* 296:2225–2229.
- Halldórsson BV, Istrail S, De La Vega FM. 2004. Optimal selection of SNP markers for disease association studies. *Hum Hered* 58:190–202.
- Hildebrandt MA, Salavaggione OE, Martin YN, Flynn HC, Jalal S, Wieben ED, Weinshilboum RM. 2005. Human *SULT1A3* pharmacogenetics: gene duplication and functional genomic studies. *Biochem Biophys Res Commun* 321:870–878.
- Horvath S, Xu X, Laird NM. 2001. The family based association test method: strategies for studying general genotype-phenotype associations. *Eur J Hum Genet* 9:301–306.
- Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM. 2004. Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genet Epidemiol* 26:61–69.
- Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RCJ, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SCL, Clayton DG, Todd JA. 2001. Haplotype tagging for identification of common disease genes. *Nat Genet* 29:233–237.
- Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs FA, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* 31:51–54.
- Ke X, Cardon LR. 2003. Efficient selective screening of haplotype tag SNPs. *Bioinformatics* 19:287–288.
- Laird NM, Horvath S, Xu X. 2000. Implementing a unified approach to family based tests of association. *Genet Epidemiol* (Suppl) 19:S36–S42.
- Nickel RG, Willadsen SA, Freidhoff LR, Huang SK, Caraballo L, Naidu RP, Levett P, Blumenthal M, Banks-Schlegel S, Bleecker E, Beaty T, Ober C, Barnes KC. 1999. Determination of Duffy genotypes in three populations of African descent using PCR and sequence-specific oligonucleotides. *Hum Immunol* 60:738–742.
- Nicodemus KK, Liu W, Chase GA, Tsai YY, Fallin MD. 2005. Comparison of type I error for multiple test corrections in large single-nucleotide polymorphism studies using principle components versus haplotype blocking algorithms. *BMC Genet* 6(Suppl 1):S78.
- Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BTN, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SPA, Cox DR. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294:1719–1723.
- Rabinowitz D, Laird N. 2000. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered* 50:227–233.
- Sebastiani P, Lazarus R, Weiss ST, Kunkel LM, Kohane IS, Ramoni MF. 2003. Minimal haplotype tagging. *Proc Natl Acad Sci USA* 100:9900–9905.
- Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC. 2002. Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the multiethnic cohort study. *Hum Hered* 55:27–36.
- Zambelli-Weiner A, Ehrlich E, Stockton ML, Grant AV, Zhang S, Levett PN, Beaty TH, Barnes KC. 2005. Evaluation of the CD14/-206 polymorphism and house dust endotoxin exposure in the Barbados Asthma Genetics Study. *J Allergy Clin Immunol* 115:1203–1209.
- Zhang K, Jin L. 2003. HaploBlockFinder: haplotype block analyses. *Bioinformatics* 19:1300–1301.