Genetic Epidemiology 30: 627–635 (2006)

A Scan Statistic for Identifying Chromosomal Patterns of SNP Association

Yan V. Sun,[1] Albert M. Levin,[1] Eric Boerwinkle,[2] Henry Robertson,[3] and Sharon L.R. Kardia[1]*

[1]*Department of Epidemiology, University of Michigan, Ann Arbor, Michigan*
[2]*Human Genetics Center, University of Texas Health Sciences Center, Houston, Texas*
[3]*Department of Biostatistics, University of Michigan, Ann Arbor, Michigan*

We have developed a single nucleotide polymorphism (SNP) association scan statistic that takes into account the complex distribution of the human genome variation in the identification of chromosomal regions with significant SNP associations. This scan statistic has wide applicability for genetic analysis, whether to identify important chromosomal regions associated with common diseases based on whole-genome SNP association studies or to identify disease susceptibility genes based on dense SNP positional candidate studies. To illustrate this method, we analyzed patterns of SNP associations on chromosome 19 in a large cohort study. Among 2,944 SNPs, we found seven regions that contained clusters of significantly associated SNPs. The average width of these regions was 35 kb with a range of 10–72 kb. We compared the scan statistic results to Fisher's product method using a sliding window approach, and detected 22 regions with significant clusters of SNP associations. The average width of these regions was 131 kb with a range of 10.1–615 kb. Given that the distances between SNPs are not taken into consideration in the sliding window approach, it is likely that a large fraction of these regions represents false positives. However, all seven regions detected by the scan statistic were also detected by the sliding window approach. The linkage disequilibrium (LD) patterns within the seven regions were highly variable indicating that the clusters of SNP associations were not due to LD alone. The scan statistic developed here can be used to make gene-based or region-based SNP inferences about disease association. *Genet. Epidemiol.* 30:627–635, 2006. © 2006 Wiley-Liss, Inc.

Key words: Poisson process; single nucleotide polymorphism; disease association; genome-wide association

Contract grant sponsor: National Institute of Health; Contract grant numbers: HL54457; HL68737.
*Correspondence to: Sharon L.R. Kardia, Ph.D., Department of Epidemiology, School of Public Health, University of Michigan, 611 Church Street, #246, Ann Arbor, MI 48104-3028. E-mail: skardia@umich.edu
Received 28 November 2005; Revised 30 January 2006; Accepted 3 June 2006
Published online 20 July 2006 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20173

INTRODUCTION

High-throughput genotyping capabilities and technologies ranging from multiplexing assays to the Affymetrix gene chips [Matsuzaki et al., 2004] are changing the face of genetic epidemiological studies focusing on identifying new genes for common chronic diseases. For example, it is now possible to do genome-wide association studies in epidemiological samples. With the growing emphasis on dense single nucleotide polymorphism (SNP) and genome-wide association studies, there is an increasing need for new analytical methods to identify significant chromosomal regions with genetic influence on common disease risk.

Several different analytical methods have been proposed for dense SNP association studies [Cheng et al., 2003; Meng et al., 2003; Lin and Altman, 2004; Neale and Sham, 2004], but most do not take into account the biological organization of the SNP data on chromosomes or only consider the linkage disequilibruim (LD) structure to define haplotype blocks. The haplotype association methods come the closest to exploiting the chromosomal position of the genetic data when applied as a moving window analysis across a chromosome [Cheng et al., 2003; Meng et al., 2003]. In situations where haplotypes are applicable, they have been shown to outperform single-locus testing [Zhang et al., 2002; Wu et al., 2004]. However, proposed moving window approaches have some of the same data-driven weaknesses as the moving window approaches used for identifying chromosomal patterns of gene expression. These weaknesses include subjective selection of the moving window size, lack of a standard

methodology for rigorously testing applicable hypotheses of association, and haplotype estimation issues in regions of low LD.

In recent years, scan statistics have been receiving more attention as a method for genomic analysis. The scan statistics arise naturally in the scanning of time and space, looking for clusters of events. In particular, scan statistics have been used within the field of molecular biology to identify chromosomal regions harboring a greater than expected number of restriction sites [Karlin and Macken, 1991] or clusters of transcription factor binding sites [Wagner, 1997, 1999; Su et al., 2001] potentially indicating groups of co-regulated genes. Hoh [Hoh and Ott, 2000] proposed the use of a simple scan statistic for linkage studies to refine the search for new genes. While each new application sheds light on how scan statistics can uncover features of the genome, the current methodologies often rely on the unrealistic assumption of uniform gene density and uniform spacing between genetic elements.

We have developed a scan statistic that takes into account the complex landscape of the human genome in the identification of significant patterns of SNP associations in their chromosomal context. This scan statistic has wide applicability for genomic analysis, including the ability to detect potential regions of DNA amplification in tumors from gene expression profiles [Levin et al., 2005]. In this paper, we extend this scan statistic to identify new genes for common disease in genome-wide association studies or dense SNP mapping to follow-up linkage peaks. Current methods for performing genome-wide associations or dense SNP association mapping of positional candidate SNPs, such as a haplotype window approach, utilize the order of SNPs on a chromosome. However, these methods tacitly assume that the distances between SNPs are fixed across the chromosome. For these reasons, we have developed a scan statistic that incorporates variation in SNP distances in an effort to identify chromosomal regions with clustered SNP association patterns. Different from haplotype tests, this scan statistic is based on single-locus test and is likely to be most useful when haplotype tests are not appropriate or possible (e.g. low LD between markers or widely spread SNPs). In an example application, we demonstrate the utility of the new scan statistic for identifying chromosomal regions with clustered significant SNPs that are associated with a common chronic disease.

METHODS

As a proof of concept, we used a sample of 1,041 unrelated Caucasians with clinically diagnosed disease status. They were genotyped for a total of 2,944 SNPs on chromosome 19 (Boerwinkle E, unpublished data). Each SNP was mapped to the chromosome 19 reference sequence of the human genome using the UCSC Golden Path databases and the chromosomal positions were recorded for statistical analysis. Single SNP associations with disease status were assessed using a χ^2 test when all three genotypes were observed and the least frequent genotype observation was more than 5. In situations where these criteria were not met, Fisher's exact test was used to test the association. The p-values from single SNP association tests were recorded and used to compare results between a sliding window method and the scan statistic method (as described below). The R^2 measure of pair-wise SNP LD was calculated using SNP Assistant version 1.0.9 (BioData, Ltd.).

SIMPLE SLIDING WINDOW APPROACH

Neal and Sham [2004] recently proposed using Fisher's [1932] product method to make gene-based inferences in genetic association studies. Using a simple sliding window approach, this statistic can also be used to identify chromosomal regions with significant SNP effects. The test statistic combines p-values from multiple independent tests of the same hypothesis and involves the χ^2 calculation

$$\chi^2_{2m} = -2 \times \sum_{i=1}^{m} \ln(p_i) \quad (1)$$

where m is the number of p-values, p_i is the p-value of the ith hypothesis test, and the χ^2 distribution has $2m$ degrees of freedom (df), under the null hypothesis. In this study, we investigated sliding windows of size 3, 4, and 5 consecutive SNPs and then merged results across adjacent windows to detect larger regions of SNP association. For example, two overlapped regions were detected in the first scan, R1 = (SNP1, SNP2, SNP3, SNP4, SNP5), and R2 = (SNP3, SNP4, SNP5, SNP6). In the merging step, if the extended region, R = (SNP1, SNP2, SNP3, SNP4, SNP5, SNP6), has a p-value lower than the α level of 0.01, region R with six SNPs is reported as the final result.

SCAN STATISTIC METHODOLOGY

Scan statistic methodologies have used two common approximations to the distribution of

events over time or space in order to develop statistical theory which can be used to identify clusters. These two common approximations fall into two categories, depending on whether or not one considers the total number of events (i.e. in our case the number of SNPs on a chromosome) to be a fixed or a random quantity. For the fixed or ''conditional'' case (conditioning on a fixed N), the most common method [Wallenstein and Neff, 1987] for computing scan statistics does so as a function of binomial probabilities assuming events occur randomly along a uniform distribution. In the non-fixed or ''unconditional'' case, the total number of events is taken to be a random variable, and similar to the conditional case, the positions of events are expected to occur at random. A Poisson process model is most often used [Conover, 1979] for the unconditional case because it conforms well to the notion of a random number of events distributed across some time/space dimension. The waiting times or distances between events in a Poisson process are independent and identically distributed exponential random variables. Given that the number of SNPs in a population or sample is better approximated as a random variable than a fixed quantity and that the distribution of SNPs is more likely to be exponentially distributed rather than uniformly distributed on chromosomes, a Poisson approximation model was used to develop the scan statistic method presented here. We acknowledge that the distribution of SNPs used in any genetic analysis results from a mixture of processes (e.g. evolutionary forces, available SNP information, investigator-driven SNP selection routines, ability to genotype, minor allele frequency detectable in the sample) that may not conform to the Poisson process assumptions. It is important that the distribution of distances between SNPs in any analysis using this scan statistic method be examined before analysis and the distances transformed (if necessary) appropriately so that the underlying distribution of distances is approximately exponentially distributed.

The scan statistic developed here expands upon the method developed by Wagner [1999] and is influenced by the landmark paper in this area of scan statistic research [Dembo and Karlin, 1992]. The scan statistic for identifying SNP association clusters uses two data types—the base pair SNP position and the p-value of SNP-disease association. To detect a regional SNP association with a disease, statistical evidence of both a clustering of SNP locations and a clustering of low p-values within that cluster of SNPs is required. For example, SNPs that have low p-values but span a large genomic region are not as likely to be detected as a SNP association cluster by our scan statistic.

To adequately describe the method, we start with a description of the simple Poisson process underlying the scan statistic. A simple Poisson process is a counting process denoted by $\{N(t), t \geq 0\}$, where $N(t)$ is a count of the number of events that occur in some time or space of length t. To identify clusters of SNPs on a chromosome, the occurrence of a SNP on the chromosome is considered to be the event of interest. Accordingly, let $N(t)$ be the number of SNPs which occur over a given base pair length t on a chromosome described by a single parameter, λ_g, which is the rate of occurrence of SNPs over a distance of t base pairs (i.e. $N(t)/t$). The expected number of SNPs, E[$N(t)$], over a region of t base pairs is equal to $\lambda_g{}^*t$. For a given group of $N(t)$ SNPs, we can test the null hypothesis that the $N(t) = \mathrm{E}[N(t)]$ against the alternative that $N(t) > \mathrm{E}[N(t)]$. Rejecting the null in favor of the alternative would be consistent with a particular group of SNPs being identified as a SNP cluster on a chromosome.

Reformulating this Poisson process in terms of a scan statistic, we consider the r distances between the $N(t) = r+1$ SNPs, where r represents the number of intervals between $N(t)$ SNPs. Let X_i represent the position of the ith SNP on the chromosome, then the distance between SNP i and SNP $i+1$ can be described as $Y_i = X_{i+1} - X_i$. For a group of $r+1$ SNPs, the distance from SNP i to SNP $i+r$ may be expressed as the sum of the r distances, $S_{i,r}$, between these $r+1$ SNPs such that $S_{i,r} = \sum_i^{i+r-1} Y_i$. Since $\{N(t), t \geq 0\}$ is a simple Poisson process, the Y_i's are independent identically distributed exponential random variables with parameter λ_g. Also, $S_{i,r}$ is distributed as a gamma random variable with rate parameter λ_g, shape parameter r, and density as follows:

$$f(s_{i,r}) = \frac{\lambda_g}{\Gamma(r)} (\lambda_g s_{i,r})^{r-1} \mathrm{e}^{-\lambda_g s_{i,r}}$$

where the gamma function $\Gamma(r) = (r-1)!$. This density function can be used to estimate the probability of a cluster of $r+1$ SNPs over a base pair distance of $S_{i,r}$. Specifically, the probability of observing a cluster of $r+1$ SNPs over a base pair distance as short or shorter than the observed

value of $S_{i,r}$ is computed from the density $f(S_{i,r})$ as follows:

$$P(S_{i,r} < s_{i,r}) = \frac{\lambda_g}{\Gamma(r)} \int_0^{s_{i,r}} (\lambda_g s_{i,r})^{r-1} e^{-\lambda_g s_{i,r}} \, ds_{i,r}. \quad (2)$$

The statistical significance of the calculated probability can be evaluated by comparing it to some previously determined α level (e.g. $\alpha = 0.01$). If the observed probability is smaller than this selected α level, then the group of SNPs is identified as a cluster of SNPs not likely to have occurred by chance alone.

Although a simple scan statistic using a Poisson process model is sufficient to detect clusters of SNPs on a chromosome, a compound Poisson process model is necessary to further incorporate SNP-disease association information. Briefly, the compound Poisson process model involves partitioning the simple Poisson process model $\{N(t), t \geq 0\}$ used to characterize the distances between SNPs into two independent Poisson processes: one for SNPs exceeding a particular SNP-disease association threshold, $\{N_1(t), t \geq 0\}$, and a second for those that do not, $\{N_0(t), t \geq 0\}$.

The justification for considering the two Poisson processes, $\{N_1(t), t \geq 0\}$ and $\{N_0(t), t \geq 0\}$ as independent arises from the following consideration. For tag SNPs that are chosen to be uncorrelated from other tag SNPs on a chromosome, their association with disease is also expected to be independent. Although epistasis among SNPs may create dependencies among SNP tests, these are not likely to be dependent on distance and hence not affect inferences about clustering. Therefore, the distribution of significant SNPs on a chromosome can be considered to be independent of the distribution of non-significant SNPs. In cases where SNPs are correlated in their frequency, the tests of association will be correlated as a function of LD and *p*-values of association will need to be decorrelated (see discussion for example) or correlated SNPs removed in order for the condition of independence to hold. We note that violation of the independence assumption will simply result in regions being judged significant based on its correlation structure.

Let P_i be the *p*-value for the *i*th SNP ordered on the chromosome ($i = 1$ to m SNPs). Based on the P_i's, an indicator variable I_i is defined to classify the *p*-value for a particular SNP i as significant or not—for example, $I_i = 1$ if $P_i < 0.1$ and "0" otherwise. Using the *p*-value from the χ^2 test or Fisher's exact test for SNP-disease association, rather than the test statistic value, ensures that all SNP comparisons occur over the same distribution and facilitate the definition of a common threshold value equally applicable to all SNPs. Because the Poisson process $\{N_1(t), t \geq 0\}$ is a subset of the original process, we can say that SNPs with significant *p*-values occur at a portion of the rate λ_g. In other words, SNPs with significant *p*-values occur at a rate equal to $\lambda_1 = \lambda_g p_1$, where p_1 is the probability that a SNP's *p*-value is below the specified threshold. Setting the threshold at 0.1, for example, $p_1 = P(I_i = 1) = 0.1$. Likewise, SNPs without significant *p*-values occur at a rate equal to $\lambda_0 = \lambda_g(1-p_1)$ such that $\lambda_1 + \lambda_0 = \lambda_g(p_1) + \lambda_g(1-p_1) = \lambda_g$. Based on these two independent Poisson processes, we define the compound Poisson process, $\{U(t), t \geq 0\}$, for identifying regions of significant SNP association clusters. *U(t)* is the sum of the independent and identically distributed I_i as follows: $U(t) = \sum_i^{i+N(t)-1} I_i$. Therefore, *U(t)* counts the number of SNPs with *p*-values below the set threshold over a base pair distance t containing a total of *N(t)* SNPs. The same type of formulation could be applied to identify clusters of non-significant *p*-values.

In the original scan statistic described above, $S_{i,r} = \sum_i^{i+r-1} Y_i$ which equals the total distance between SNP i to SNP $i+r$. Therefore, the base pair width of an identified cluster of significant SNP associations between SNP i and SNP $i+r$ is then represented by $S_{i,r,k} = S_{i,r} = \sum_i^{i+r-1} Y_i$ where k (similar to r above) is the number of intervals between *U(t)* highly significant SNPs. This $S_{i,r,k}$ statistic is used to calculate the probability of a highly associated cluster based on the gamma density as follows:

$$P(S_{i,r,k} < s_{i,r,k}) = \frac{\lambda_1}{\Gamma(k)} \int_0^{s_{i,r,k}} (\lambda_1 s_{i,r,k})^{k-1} e^{-\lambda_1 s_{i,r,k}} \, ds_{i,r,k}. \quad (3)$$

In the case where $U(t) = N(t)$ (i.e. the number of SNPs observed over a base pair distance of size t are all significant), k is equal to r and the probability calculation is virtually identical to the probability calculated using equation (2), with the exception of substituting λ_1 for λ_g.

A fundamental assumption underlying this scan statistic is that the distance between SNPs is exponentially distributed. For the 2,944 SNPs genotyped in this study, we used their chromoso-

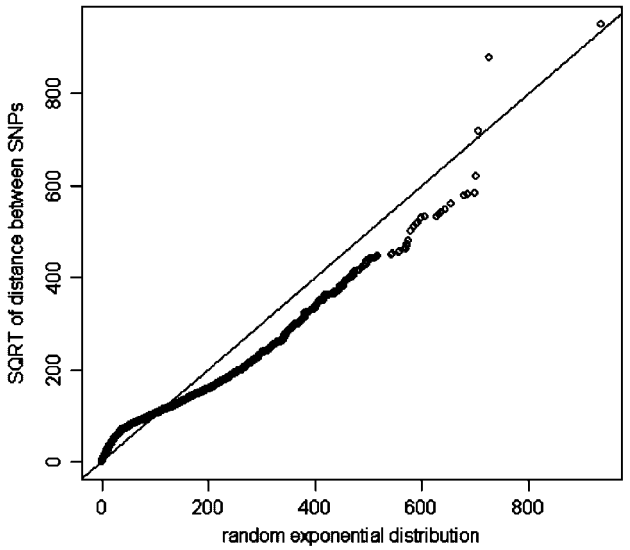

Fig. 1. QQ- plot to compare square root of distances of 2,944 human chromosome 19 SNPs to a matching exponential distribution. Square root of SNP distances was plotted as *y*-axis. A random exponential distribution was generated with same λ as the 2,943 square root of SNP distances (1/mean) and plotted as *x*-axis. R^2 was calculated by linear regression of two data sets.

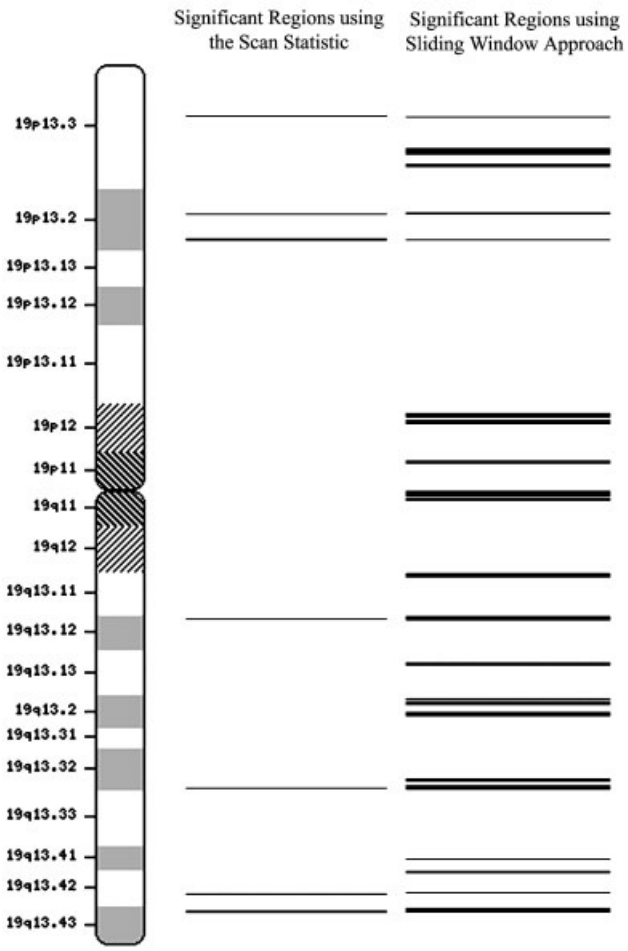

Fig. 2. A comparison of the statistically significant regions identified using the scan statistic and sliding window approaches. Seven significant regions detected by scan statistic and 22 significant regions detected by a sliding window algorithm were aligned with human chromosome 19. Region *p*-value cutoff was 0.01.

mal position to obtain all distances between neighboring SNP pairs. Raw and transformed SNP distance distributions were compared with matching random exponential distribution using the R statistical software package. In Figure 1, we display the quantile-quantile (Q-Q) plot of the SNP-SNP distances that indicate that a square root transformation of the distances fits an exponential distribution for the majority of SNPs on chromosome 19.

RESULTS

We applied our SNP association scan statistic to the 2,944 chromosome 19 SNPs using a *p*-value threshold of 0.1 to categorize each single SNP test of disease association as an event or not, and statistically significant regions were determined using the scan statistic using a regional α level of 0.01 (same as sliding window). The scan statistic detected seven significant regions containing between 3 and 6 SNPs (Fig. 2 and Table I) and ranging from 10 to 73 kb in size. The average span of these regions was about 35 kb, which is much shorter than the average span of the sliding window results. The most significant region ($p = 0.00069$) contained 6 SNPs and spanned 40 kb. LD patterns were then investigated within significant SNP association clusters between significant and non-significant SNPs in each region. In Table II, the R^2 measure of LD between pairs of SNPs is plotted to demonstrate the LD structure within each region. Since the scan statistics algorithm was only based on SNP association and chromosomal position, it was not surprising to see that the identified regions have very different LD structures. Only one of the seven regions, region 3, displayed very tight LD among significant SNPs, while the remaining six regions had a wide range of LD, indicating that the region was not identified solely because of LD.

TABLE I. Distribution of linkage disequilibrium among SNPs that were significantly associated with disease versus not significantly associated in the identified regions

Region ID	Total SNPs	significant SNPs	non-significant SNPs	Region Size (bp)	Region p-value	LD R^2 between significant SNPs	LD R^2 between significant and non-significant SNPs
1	5	4	1	22131	2.00E-03		
2	5	4	1	72991	7.08E-03		
3	4	3	1	10101	8.71E-03		
4	3	3	0	15601	7.78E-03		NA
5	5	4	1	44201	3.20E-03		
6	6	5	1	40501	6.90E-04		
7	5	4	1	38901	3.46E-03		

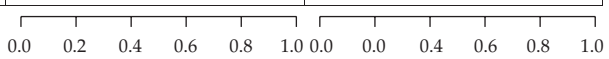

TABLE II. Chromosome 19 SNP-disease association results of sliding window scan

Region ID	Total SNPs	Significant SNPs	Non-significant SNPs	Region size (bp)	Region *p*-value
1	8	4	4	61741	0.004882
2	5	3	2	77561	0.005757
3	8	3	5	66571	0.006071
4	8	4	4	42961	0.00423
5	5	4	1	72991	0.006874
6	4	3	1	10101	0.005016
7	5	2	3	239001	0.008791
8	5	4	1	188801	0.005297
9	3	2	1	91801	0.008089
10	8	2	6	615001	0.002657
11	7	4	3	144201	0.006726
12	7	3	4	202101	0.008999
13	4	2	2	118601	0.009398
14	5	3	2	365701	0.006981
15	5	3	2	228201	0.008235
16	6	3	3	32301	0.006633
17	15	7	8	146701	6.06E-06
18	4	1	3	11201	0.008351
19	5	3	2	35901	0.005172
20	9	5	4	52801	0.000545
21	7	4	3	53401	0.008659
22	4	2	2	21901	0.009433

The sliding window approach using Fisher's product method was applied to chromosome 19 SNP data using a regional α level of 0.01. Using a fixed window size of 3 SNPs, there were 20 genomic regions detected with *p*-values lower than the above threshold. The increased window size of 4 and 5 SNPs each detected 17 significant regions. Combining the significant regions from each of the individual fixed window sizes yielded a total of 22 statistically significant regions (Table II). The number of SNPs in these regions ranged from 3 to 15. Using this method, we found nine regions that spanned greater than 100 kb. The smallest region spanned 10.1 kb and the largest region spanned 615 kb. The average span was 131 kb. The most significant region ($p = 6 \times 10^{-6}$) included 15 SNPs and spanned 147 kb. All seven regions detected by the scan statistics approach were also found to be significant using the sliding window approach (Fig. 2), despite some slight differences in region sizes.

The individual SNP *p*-values of significant regions and region sizes from the scan statistic and sliding window approaches are presented in Figure 3A and B, respectively. Among the seven scan statistic regions, there was at most one SNP with a *p*-value above 0.1 in each region. Conversely, 18 of the 22 sliding window regions

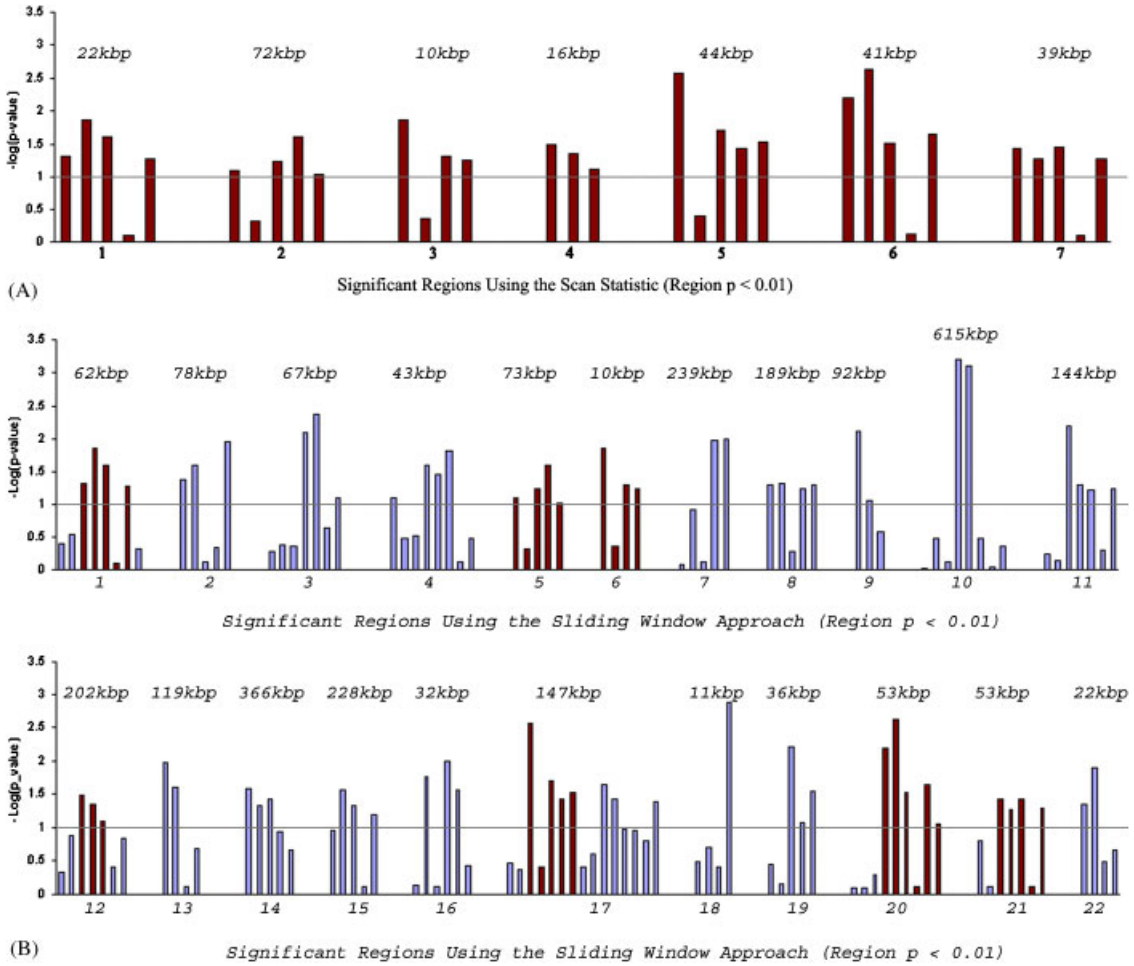

Fig. 3. Individual *p*-value and region size of all significant regions from using the scan statistic (A) and sliding window approaches (B). Common SNPs detected by both algorithms are in dark color.

contained more than one SNP with *p*-value above 0.1. In addition, only one out of four SNPs within region 18 had a significant *p*-value. This may indicate that the regions with fewer significant SNPs but more significant individual *p*-values tend to be picked up by the sliding window algorithm incorrectly. Comparing the seven overlapped regions detected by both algorithms, five of seven had longer spans in the results of the sliding window approach.

In order to test the significance of the number of detected SNP clusters, we ran 1,000 permutation tests by shuffling the SNP order on the chromosome. Among the permutation test results of the sliding window approach, the highest number of detected regions was 20. For the scan statistic, the highest number of detected regions among 1,000 permutation tests was 5. The permutation results strongly support that the detected regions from the two methods are not solely due to the random effects.

DISCUSSION

With the proliferation of SNP genotyping technologies and the availability of the human genome sequence, multiple studies suggest that high-density SNP genotype data should be used to detect genes that are associated with common diseases [Risch, 2000; Cardon and Bell, 2001]. As such, the number of methods being put forth to study genome-wide disease association is growing [Cheng et al., 2003; Meng et al., 2003; Neale and Sham, 2004]. Generally, these methods use only the relative position of SNPs and assume that the distances between the SNPs are fixed and constant within and across chromosomes. In the

context of studying the association between disease and regional SNP patterns, the scan statistic proposed here represents a significant advance over previous methods because it accounts for variation in the distance between consecutive SNPs.

A major advantage of the scan statistic method is that inferences can be made for a chromosomal region rather than being restricted to single SNP inferences or a preset moving window of SNPs. Also, it is important to reiterate that this methodology identified regions where not all SNPs in a putative region were associated. In addition, this method can be used for making gene-based inferences when multiple SNPs within a gene are measured. There are many instances in the literature where multiple SNPs are measured in a single gene but the results from single SNP associations differ [Patterson and Cardon, 2005; Swarbrick et al., 2005]. In this case, the distribution of significant SNPs may be different across populations but the same gene may be implicated because it is identified as a region of SNP association. Different patterns of SNP associations with a disease outcome in separate studies are expected because of differences in the underlying frequency and LD distribution of SNPs across studies. In general, the scan statistic offers a statistical method for making regional or gene-based inferences about disease association that is likely to be more robust across studies than single SNP association results.

The ability to incorporate the spatial features of genetic elements in the genome into a statistical analysis of chromosomal patterns of SNP effects has direct implications for dense chromosomal or genome-wide SNP association studies. In this situation, the data used in the scan statistic were the chromosomal position of the SNP and the results from the single SNP associations with the phenotype of interest (e.g. a χ^2 statistic or p-value). Because the forces of LD are correlated to their physical distance (although not monotonically or necessarily exclusive of long-range LD), closely spaced SNPs that demonstrate some level of disease association are likely to represent the same underlying functional SNP or haplotype. By incorporating the exact distance between SNPs in a region, the scan statistic proposed here can be used to identify chromosomal regions with SNP patterns of association that are unlikely to occur by chance alone and provides a statistical means to prioritize regions for more detailed molecular or haplotype investigation.

In comparison to the sliding window methods to detect chromosomal patterns of SNPs association, the scan statistic approach using the compound Poisson process formulation has the advantage of being parametric. Thus, the statistical significance of clustering in the data is determined in comparison to a theoretical null distribution rather than having to estimate it using permutation strategies where the results are data-dependent and cannot be easily compared across different data sets. There are, however, important features that must be addressed when using the scan statistic for genome-wide association or dense SNP studies. For the Poisson process model to be appropriate, the distances between events must be exponentially distributed or transformed to approximate that distribution. For dense SNP studies of positional candidate genes, it may be that each gene in the chromosomal region must be analyzed separately so that this assumption holds. In genome-wide SNP studies, this may be less of an issue.

We observed a wide range of LD across the seven regions detected by the scan statistic. However, a more precise model which incorporates the LD information would improve the utility of the scan statistic method to identify important gene regions that contain multiple SNP association. We are currently evaluating methods to incorporate correlation structure into the scan statistic methods to guard against inflation of the significance of an SNP cluster, because of the correlations among SNPs due to LD. With highly correlated SNP genotyping data, a decorrelation step should be cautiously applied to adjust the single-locus test results before running the scan statistics. Zaykin et al. [2002] suggest an approach to decorrelate the p-values by premultiplying a vector of p-values by the Cholesky factor of the vector's correlation matrix. A newly proposed method [Conneely and Boehnke, 2005] for estimating the correlation matrix among p-values provides additional computationally efficient and a potentially powerful option. Another potential approach to minimize LD effects is to apply the SNP association ranks instead of p-values in the scan statistic, because recent findings suggest that the effect of even strong LD on true association ranks is too small to be of substantial importance in genome-wide association studies [Zaykin and Zhivotovsky, 2005]. Combined with better SNP selection strategies, such as a tag-SNP approach from the HapMap project [The International HapMap consortium, 2003, 2004], LD

structure would be a less important issue for the scan statistic developed here.

Overall, this scan statistic's limitation caused by the inter-SNP correlation can be corrected by the above statistical adjustments. A final advantage of the scan statistic presented here is that it can be used with any type of hypothesis test—e.g. tests of gene-environment interaction or tests of allele frequency differences across ethnic groups. In general, the chromosomal or gene-based regions detected by the scan statistic are expected to provide a better statistical foundation from which to identify these regions and to make comparisons across studies.

REFERENCES

Cardon LR, Bell JI. 2001. Association study designs for complex diseases. Nat Rev Genet 2:91–99.

Cheng R, Ma JZ, et al. 2003. Nonparametric disequilibrium mapping of functional sites using haplotypes of multiple tightly linked single-nucleotide polymorphism markers. Genetics 164:1175–1187.

Conneely KN, Boehnke M. 2005. Combining correlated p-values in trait-SNP association studies. The American Society of Human Genetics 55th Annual Meeting, Salt Lake City, Utah. 184p.

Conover WJ, Bement TR, Iman RL. 1979. On a method of detecting clusters of possible uranium deposits. Technometrics 21: 277–283.

Dembo A, Karlin S. 1992. Poisson approximations for r-scan processes. The Annals of Applied Probability 2:329–357.

Fisher RA. 1932. Statistical methods for research workers. 4th ed. London: Oliver and Boyd.

Hoh J, Ott J. 2000. Scan statistics to scan markers for susceptibility genes. Proc Nat Acad Sci USA 97:9615–9617.

Karlin S, Macken C. 1991. Assessment of inhomogeneities in an *E. coli* physical map. Nucleic Acids Res 19:4241–4246.

Levin AM, Ghosh D, et al. 2005. A model-based scan statistic for identifying extreme chromosomal regions of gene expression in human tumors. Bioinformatics 21:2867–2874.

Lin Z, Altman RB. 2004. Finding haplotype tagging SNPs by use of principal components analysis. Am J Hum Genet 75: 850–861.

Matsuzaki H, Dong S, et al. 2004. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nat Methods 1:109–111.

Meng Z, Zaykin DV, et al. 2003. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. Am J Hum Genet 73:115–130.

Neale BM, Sham PC. 2004. The future of association studies: gene-based analysis and replication. Am J Hum Genet 75:353–362.

Patterson M, Cardon L. 2005. Replication publication. PLoS Bio 3:e327.

Risch NJ. 2000. Searching for genetic determinants in the new millennium. Nature 405:847–856.

Su X, Wallenstein S, et al. 2001. Nonoverlapping clusters: approximate distribution and application to molecular biology. Biometrics. 57:420–426.

Swarbrick MM, Waldenmaier Br, et al. 2005. Lack of support for the association between GAD2 polymorphisms and severe human Obesity. PLoS Bio 3:e315.

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

The International HapMap Consortium. 2004. Integrating ethics and science in the International HapMap Project. Nat Rev Genet 5:467–475.

Wagner A. 1997. A computational genomics approach to the identification of gene networks. Nucleic Acids Res 25:3594–3604.

Wagner A. 1999. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. Bioinformatics 15:776–784.

Wallenstein S, Neff N. 1987. An approximation for the distribution of the scan statistic. Stat Med 6:197–207.

Wu S-J, Chiang F-T, et al. 2004. Three single-nucleotide polymorphisms of the angiotensinogen gene and susceptibility to hypertension: single locus genotype vs. haplotype analysis. Physiol Genomics 17:79–86.

Zaykin DV, Zhivotovsky LA, et al. 2002. Truncated product method for combining P-values. Genet Epidemiol 22:170–185.

Zaykin DV, Zhivotovsky LA. 2005. Ranks of genuine associations in whole-genome scans. Genet 171:813–823.

Zhang K, Calabrese P, et al. 2002. Haplotype block structure and its applications to association studies: power and study designs. Am J Hum Genet 71:1386–1394.