Genome analysis

ChromoScan: a scan statistic application for identifying chromosomal regions in genomic studies

Yan V. Sun*, Douglas M. Jacobsen and Sharon L. R. Kardia
Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI 48104-3028, USA
Received on August 24, 2006; revised on September 28, 2006; accepted on September 30, 2006
Advance Access publication October 10, 2006
Associate Editor: Keith A Crandall

ABSTRACT
Summary: ChromoScan is an implementation of a genome-based scan statistic that detects genomic regions, which are statistically significant for targeted measurements, such as genetic associations with disease, gene expression profiles, DNA copy number variations, as well as other genome-based measurements. A Java graphic user interface (GUI) is provided to allow users to select appropriate data transformations and thresholds for defining the significant events.
Availability: ChromoScan is freely available from http://www.epidkardia.sph.umich.edu/software/chromoscan/
Contact: yansun@umich.edu

1 INTRODUCTION
High-throughput genomic technologies ranging from gene expression microarrays, single nucleotide polymorphism (SNP) genotyping, to array-based comparative genomic hybridization (array CGH) are changing the face of biomedical studies by focusing attention on genomic, rather than single gene, features of common human diseases. To identify the regional effects of thousands of genomic markers, the naive moving window approach has been widely adapted for such analyses (Cheng et al., 2003; Meng et al., 2003). However, the complex landscapes of the human genome are not evenly distributed in its genomic features, such as human genes (Levin et al., 2005) and SNPs (Sun et al., 2006) and bias is introduced by assuming this uniformity in a moving window approach (Wagner, 1997).

Scan statistics have traditionally been used to scan both time and space for evidence of significant clusters of events. Recently, they have been receiving more attention as a method for genomic analysis. 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 (Su et al., 2001; Wagner, 1999) 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. Recent studies have extended the utilities of scan statistics to the analyses of genome-wide gene expression (Levin et al., 2005) and SNP association (Sun et al., 2006) by incorporating distance between genomic elements into the identification of significant genomic regions.

Using a compound Poisson process, the scan statistic methodology presented here takes into account the complex distribution of genome variation in the identification of chromosomal regions with significant clusters of SNP-disease associations. However, this scan statistic has wide applicability to all developing technologies for genome-wide measurements including proteomics, interference RNAs and toxicogenomic studies. It can also be used to make gene-based or region-based inferences for targeted research hypotheses.

2 IMPLEMENTATION
ChromoScan is implemented in Java and will run on all operating systems with the proper version of Java Virtual Machine (JVM) installed. The interface provides a graphical view to allow users to input any genomic measure and its base pair position check data distributions and to select appropriate parameters and data transformations to meet the assumptions underlying the scan statistic. It also provides an optional permutation test functionality to help users to evaluate the probability of detecting a certain number of significant regions. Users are also able to visualize the results and export them into character delimited text and high-resolution image files.

For example, to identify regions with evidence of clustering of SNP-disease associations there are two types of inputs—the base pair position of the SNP and the P-value of the single SNP association with disease (Sun et al., 2006). Statistical evidence of both a clustering of SNP locations and a clustering of low P-values within that cluster are required for a region to be identified as significant.

Brieﬂy, for any genomic marker or measure, consider the r distances between the N(t) = r + 1 markers, where N(t) is a count of the number of events that occur in a space of length t. Let Xi represent the position of the i th marker on the chromosome, then the distance between marker i and marker i + 1 can be described as Yi = Xi+1 − Xi. For r + 1 ordered markers, the distance from marker i to marker i + r can be expressed as Si,r = ∑i+r−1 t=i Yi. Since, {N(t), t ≥ 0} is a simple Poisson process, the Yi’s are independent, identically distributed, exponential random variables with parameter λg. Also, Si,r is distributed as a gamma random variable with rate parameter λg, shape parameter r and the probability of observing a cluster of r + 1 markers over a base pair distance shorter than the observed value of si,r is computed as follows:

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

where, the gamma function \( \Gamma(r) = (r - 1)! \). If the observed probability is smaller than a pre-selected \( \alpha \) level (e.g. \( \alpha = 0.01 \)), then the
group of markers is identified as a cluster of markers not likely to have occurred by chance alone. False positives due to multiple testing will affect inferences based on the scan statistic. Adjusting for multiple testing or using methods, such as cross-validation will increase the possibility of detecting functionally important regions. To incorporate the markers’ effects or measured level, a compound Poisson process model is used to partition the simple Poisson process model described earlier into two independent Poisson processes: one for markers exceeding a particular threshold, \( \{N_i(t), t \geq 0\} \), and a second for those that do not, \( \{N_0(t), t \geq 0\} \). Let \( P_i \) be the \( P \)-value for the \( i \)th marker. Then, based on the user defined significance threshold, an indicator variable \( I_i \) is defined to classify whether \( P_i \) is significant or not—e.g., \( I_i = 1 \) if \( P_i < 0.1 \) and ‘0’ otherwise. Because the significant Poisson process \( \{N_i(t), t \geq 0\} \) is a subset of the original process, markers with significant \( P \)-values occur at rate \( \lambda_1 \), which is a portion of the rate \( \lambda_1 - \lambda_1 = \lambda_1 p_i \), where \( p_i \) is the probability that a marker’s \( P \)-value is below the specified threshold. Likewise, markers without significant \( P \)-values occur at a rate \( \lambda_0 = \lambda_1 (1 - p_i) \). The compound Poisson process \( \{U(t), t \geq 0\} \), for identifying regions of significant marker association clusters is a function of these two Poisson processes, where \( U(t) \) is the sum of the independent and identically distributed \( I_i \) as follows:

\[ U(t) = \sum_{i=1}^{\text{(N(t)-1)}} I_i. \]

Therefore, \( U(t) \) counts the number of significant markers over a base pair distance \( t \) containing a total of \( N(t) \) markers. The base pair width of an identified cluster is then represented by \( S_{i,r,k} = S_{i,r} = \sum_{i=1}^{\text{(N(t)-1)}} Y_i \), where \( k \) is the number of intervals between \( U(t) \) significant markers. The probability of a cluster is calculated based on the gamma density as follows:

\[ P(S_{i,r,k} < S_{i,r,k}) = \frac{\lambda}{\Gamma(k)} \int_0^{S_{i,r,k}} (\lambda 1_{S_{i,r,k}})^{k-1} e^{-\lambda 1_{S_{i,r,k}}} dS_{i,r,k}. \]

In this application, sliding windows of size 3–10 consecutive markers are typically examined. Results are merged automatically across adjacent windows to delineate larger regions as long as the \( P \)-value for the joint region is less than the user defined level. To test the significance of the number of detected clusters, a permutation test randomly shuffles the marker order on the chromosome. A histogram of the number of detected regions for each permutation is plotted along with the actual number of detected regions and a \( P \)-value is estimated. On a 3 GHz Dual Core Pentium 4 desktop with 1 GB memory, a genome scan of 500 K Affymetrix chip results takes less than 15 minutes including all operations to read the data and plot the figures. In fact, the majority of the time is spent reading the input file (distances, \( P \)-values) and drawing figures for examining the distributional assumption.

3 DISCUSSION AND CONCLUSION

In order to make the scan statistic algorithm available to general biomedical researchers, we developed the GUI interface and plotting tools in Java language. This application can be run under operating systems, such as Windows, Linux and Unix with compatible version of Java Runtime Machine (JVM) installed. ChromoScan offers step-by-step interface and distribution preview to guarantee proper file input and appropriate parameter selection for the scan statistic. The quantile–quantile (QQ) plot, which compares the marker distance distribution with the exponential distribution, provides a convenient and critical tool to select the correct data transformation to meet the underlying assumption of exponentially distributed distances between markers. If the raw distances violate the exponential distribution assumption of the scan statistic, users can use a QQ plot of the marker distances against the theoretical exponential distribution with matching parameters to choose an appropriate transformation. Five transformations—square root, reciprocal, natural log, negative log and a free form of power transformation—are available in ChromoScan for users to choose. We have downloaded the tagSNPs from the HapMap project to study their distance distribution. After appropriate transformation, we have confirmed that the transformed distances within each
chromosome follow an exponential distribution (Fig. 1C shows chromosome 22 as an example).

Although ChromoScan is a flexible tool to scan genome-wide features, some limitations need to be considered. In the case of high density SNP genotyping data, the inter-marker correlations need to be cautiously handled to make appropriate regional inferences (Sun et al., 2006). Users may also select only the tagSNPs to run the scan statistic to minimize the inter-marker correlation effect.

ChromoScan creates the opportunity for general researchers to apply scan statistics on various types of genomic data, such as the ongoing genome-wide association studies. We expect that ChromoScan will have wide applicability in such genome-wide association studies to identify statistically significant regions associated with common chronic diseases.

ACKNOWLEDGEMENTS

This work was supported by National Institute of Health grant HL54457 and HL68737.

Conflict of Interest: none declared.

REFERENCES


