Published in final edited form as:

Comput Stat Data Anal. 2009 March 15; 53(5): 1794–1801. doi:10.1016/j.csda.2008.04.013.

A Fast Implementation of a Scan Statistic for Identifying Chromosomal Patterns of Genome Wide Association Studies

Yan V. Sun¹, Douglas M. Jacobsen¹, Stephen T. Turner², Eric Boerwinkle³, and Sharon L.R. Kardia¹

¹ Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan

² Division of Nephrology and Hypertension, Mayo Clinic, Rochester, Minnesota

³ Human Genetics Center, University of Texas Health Sciences Center, Houston, Texas

Abstract

In order to take into account the complex genomic distribution of SNP variations when identifying chromosomal regions with significant SNP effects, a single nucleotide polymorphism (SNP) association scan statistic was developed. To address the computational needs of genome wide association (GWA) studies, a fast Java application, which combines single-locus SNP tests and a scan statistic for identifying chromosomal regions with significant clusters of significant SNP effects, was developed and implemented. To illustrate this application, SNP associations were analyzed in a pharmacogenomic study of the blood pressure lowering effect of thiazide-diuretics (N=195) using the Affymetrix Human Mapping 100K Set. 55,335 tagSNPs (pair-wise linkage disequilibrium $R^2 < 0.5$) were selected to reduce the frequency correlation between SNPs. A typical workstation can complete the whole genome scan including 10,000 permutation tests within 3 hours. The most significant regions locate on chromosome 3, 6, 13 and 16, two of which contain candidate genes that may be involved in the underlying drug response mechanism. The computational performance of ChromoScan-GWA and its scalability were tested with up to 1,000,000 SNPs and up to 4,000 subjects. Using 10,000 permutations, the computation time grew linearly in these datasets. This scan statistic application provides a robust statistical and computational foundation for identifying genomic regions associated with disease and provides a method to compare GWA results even across different platforms.

Keywords

Scan Statistic; Genome Wide Association; Single Nucleotide Polymorphism; Distance Distribution; ChromoScan

Introduction

With the growing emphasis on genome-wide association studies for finding new genomic loci for common chronic disease [1–3], there is an increasing need for new analytical methods to identify significant chromosomal regions with genetic influence on these diseases. We

Corresponding author: Yan V. Sun, Ph.D., University of Michigan, Department of Epidemiology, School of Public Health, 109 Observatory, #4605, Ann Arbor, MI 48109, Phone: (734) 615-6279, Fax: (734) 764-1357, Email address: yansun@umich.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

developed and implemented a scan statistic and a permutation procedure to identify disease-associated genomic regions which takes as input the SNP genotypes, qualitative outcomes and chromosomal locations of all SNPs.

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 [4] or clusters of transcription factor binding sites [5–7] potentially indicating groups of co-regulated genes. Hoh [8] proposed the use of a scan statistic for linkage studies to refine the search for new genes.

We have developed a scan statistic that takes into account the complex landscape of the human genome in the identification of significant patterns of gene expression profiles [9] and SNP associations [10] in their chromosomal context. The SNP scan statistic has wide applicability for genomic analysis, including the ability to detect potential regions of copy number variation (CNV) and regions of significant gene-environment interactions. For general applications of the scan statistic, we implemented a Java graphic user interface, ChromoScan, to allow users to select appropriate data transformation and thresholds for defining the significant genomic events [11].

Although in theory, a fixed window size can result in a global p-value without any need for multiple testing correction [8], a correction for multiple testing is necessary for real genomic data which contain various window sizes. In this paper, we implemented a fast permutation test strategy correcting for the multiple testing bias on the highdimensional genome-wide association (GWA) analysis. A new version of ChromoScan, ChromoScan-GWA, utilizes genome-wide SNP data and the chromosomal location of SNPs, and integrates single SNP tests (e.g. chi-square test for case-control data) to allow evaluations of the regional inference of clustered SNP association patterns. ChromoScan-GWA is available at: <http://www.epidkardia.sph.umich.edu/software/chromoscan/>

To illustrate the application, we analyzed patterns of SNP association in a genome-wide SNP dataset. In a genome-wide screen for genes influencing blood pressure response to thiazide diuretics, 195 (98 high responders and 97 low responders) independent Caucasians were genotyped at a total of 112,990 SNP loci. Single SNP association with drug response was assessed by chi-square test. P-values from single SNP association tests were recorded and used for the scan statistic analyses. To calculate the empirical p-values adjusted for multiple testing, the regional p-values from the scan statistic were compared to a distribution of p-values from 10,000 permutation tests. In order to demonstrate the utility of ChromoScan-GWA in large GWA studies, we also tested its scalability by running a series of simulation datasets with up to 1,000,000 SNPs and up to 4,000 subjects in 10,000 permutation tests.

Methods

The Scan Statistic for Identifying Significant SNP Clusters

In order to develop statistical theory of identifying clusters of events, scan statistic methodologies have used two common approximations to the distribution of events over time or space, depending on whether one considers the total number of events (i.e. in our case the number of SNPs on a chromosome) to be a fixed (“conditional”) or a non-fixed (“unconditional”) quantity. In the non-fixed or “unconditional” case, the total number of events is taken to be a random variable. A Poisson process model is most often used [12] 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 on a chromosome 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. 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 SNPs that occur in a chromosomal region of length t , and is described by a single parameter, λ_g , which is the rate of occurrence of SNPs over a distance of t base pairs (i.e. $N(t)/t$).

Let P_i be the p-value from a chi-square test of independence 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 in this application, $I_i = 1$ if $P_i < 0.05$ and otherwise, $I_i = 0$. Using the p-value, rather than the test statistic value, ensures that all SNP comparisons occur over the same distribution and facilitates 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 0.05, for example, $p_1 = P(I_i = 1) = 0.05$. 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:
$$U(t) = \sum_{i=1}^{i+N(t)-1} I_i$$
 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. The base pair width of an identified cluster is then

represented by $S_{i,r,k} = S_{i,r} = \sum_{i=1}^{i+r-1} Y_i$, and is used to calculate the probability of a highly associated cluster based on the gamma density as follows:

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

where k (similar to r above) is the number of intervals between $U(t)$ highly associated SNPs. 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 .

Permutation Test

High-dimensional GWA data involve 100,000's of SNPs which cause serious concern for false positive results due to multiple testing. Methods have been developed for correcting multiple testing biases in single SNP and haplotype moving window analysis in GWA studies [13,14]. For the scan statistic application presented here, we investigated the impact of multiple testing and implemented a new permutation test to estimate the empirical p-values in the context of a large amount of single SNP tests. Our previous permutation test shuffled p-values across locations rather than repeating the entire analyses on each permuted dataset.

The new permutation testing strategy works by shuffling outcomes with respect to the marker genotypes for each subject and then re-computing significant regions a user-specified number (e.g. 10,000 in this study) of iterations [15]. The empirical p-value for each region is determined by examining the distribution of the p-values associated with clustered SNP effects within the

permutation data. For each iteration of the permutation test the following steps are performed. First, the outcome values are shuffled with respect to the marker genotypes, and p-values associated with the shuffled outcomes are derived for each SNP from chi-square tests of independence. Second, the most significant chromosomal regions are identified using the scan statistic algorithm on each permuted dataset. Permutation region significance is determined with a more relaxed alpha cutoff to account for the increased p-value variation caused by shuffling. Specifically, we used $\alpha=0.01$ for initial region discovery and $\alpha=0.1$ for permutation region discovery. Third, the significant regions identified in a permutation dataset were ordered by p-value, and the best n regions were retained for subsequent analysis, where n is the number of regions initially discovered (prior to permutation testing). If only k ($k < n$) significant regions are identified in a particular iteration of the permutation procedure, then p-values = 1.0 are used for the remaining $n-k$ p-values in that permutations distribution. Because the regional α for the permutation was reset to a more liberal level ($\alpha=0.1$) than the original scan ($\alpha=0.01$), the total number of significant regions (k) in each permutation is unlikely to be less than n . We implemented this step to prevent potential errors in extreme cases. Since the complete scan has to be performed to rank the most significant regions in each iteration, the number of the selected top regions (i.e. n) does not affect the computation of the scan statistic. After all iterations were completed, the total list of significant regions identified in all 10,000 permutation datasets was sorted by p-value. The empirical p-value for each region was calculated by computing the ratio of the number of permutation regions with more significant p-values (than the observed region's p-value) against the total number of permuted regions.

Computational Implementation

The new ChromoScan-GWA software was written in Java, using the 1.5 JDK and extends the ChromoScan scan statistic framework [11]. ChromoScan-GWA computes significant genomic regions associated with a dichotomous outcome. The significant regions are discovered, one chromosome at a time, using the ChromoScan scan statistical algorithm. Briefly, the scan statistic identifies significant regions by finding groups of neighboring significant markers accounting for the genomic distribution of the SNP, followed by a coalescence step where significance of each region is evaluated based on a p-value and user-defined significance cutoff (i.e. α level). The sliding windows of size 3–10 consecutive markers are examined for each scan, the same as the original ChromoScan software [11]. The regions with significant p-values are merged automatically across adjacent windows to delineate the non-overlapped regions as long as the joint region has the lowest regional p-value. Finally, a permutation test was implemented to correct multiple testing bias of genome-wide SNP data. In ChromoScan-GWA, individual marker significance is assessed using a Chi-Square test of independence. P-values are computed from the Chi-Square test statistic by the Colt statistical library [16].

GWA data analysis

Data

The sample consisted of 195 non-Hispanic whites with hypertension who were selected from opposite tertiles of the race and sex-specific distributions of age-adjusted drug response. The sample of 195 non-Hispanic whites consisted of 98 “high” responders (age = 47.1 ± 6.1 years, 51.5 % women) and 97 “low” responders (age = 47.4 ± 6.5 years, 51.5% women).

Of the 116,201 SNPs on the Affymetrix GeneChip® Human Mapping 100K array set, the number of SNPs suitable for statistical analysis was reduced to 94,609 (81.4% of total) by the exclusion of SNPs for the following reasons (in order): on the X-chromosome, monomorphic, relative frequency of “no-call” >20% across samples, and deviation from Hardy-Weinberg equilibrium at $P < 0.001$.

ChromoScan Analysis

Using the visualization tool implemented in ChromoScan [11], we verified that the tagSNPs were approximately distributed exponentially on the chromosomes after appropriate power transformation (between the power of 0.6 to 0.7). The power transformed distances were used in the following SNP scan statistic and permutation tests.

To reduce the correlation among the SNPs [10] and improve the computation performance, 5,335 tagSNPs were picked by the threshold of pair-wise LD R^2 less than 0.5 [17] from 22 chromosomes. The numbers of tagSNPs from each chromosome are summarized in Table 1. Using the chi-square test, we estimated the drug response association of each tagSNP and applied the SNP scan statistic.

We identified 50 regions significantly associated with the drug response at a regional p-value cutoff of 0.01. Seventeen out of twenty two chromosomes contained at least one associated region. There were no regions identified on chromosome 14, 15, 17, 19 and 21. Chromosome 7 had the highest number of significant regions (i.e. eight significant regions). The size of the 50 identified regions ranged from 341 bp (3 tagSNPs) to 542,553 bp (11 tagSNPs) with a mean size of about 137 kbp and a median of 84.5 kbp.

Permutation tests

We calculated the empirical p-values for the 50 identified regions by comparing the original p-values to the lowest p-values from the 10,000 permutation tests (shuffling the drug response, re-calculating the chi-square p-values, re-calculating the regional p-values). In Table 2, we list the empirical p-values based on the 10,000 permutation tests. Four regions remained significant with empirical p-values less than 0.05. The scan statistic regional p-values of the four regions were all less than 10^{-4} .

By comparing the genomic locations of the four regions, we found that two out of four overlap known genes. Region 12 spanned 142,911 bp on chromosome 3. There are 10 tagSNPs in the region and 5 of them had chi-square p-values less than 0.05. This region covers the 3' flanking region, promoter region and the last two exons of succinate-CoA ligase, GDP-forming, beta subunit (SUCLG2). Region 47 overlapped with cadherin 13, H-cadherin (CDH13) on chromosome 16. This gene is a member of the cadherin superfamily. This cadherin is a putative mediator of cell-cell interaction in the heart and may act as a negative regulator of neural cell growth. The gene locus is hypermethylated or deleted in breast, ovarian and lung cancers. For the significant genomic region 12 overlapping SUCLG2, the single SNP tests and inter-SNP correlation (LD) are summarized in Figure 1.

Computational Performance

In order to demonstrate the computational scalability of ChormoScan-GWA in a large GWA dataset, we created a series of datasets by replicating the existing real dataset. For each of the three sample sizes, $n=200$, $n=2,000$ and $n=4,000$, datasets of 50K, 100K, 250K, 500K and 1M SNPs were produced. The computation time of these datasets were summarized in Figure 2 based on 10,000 permutation tests on a Linux server with 1.6 GHz dual quad-core CPUs (a single core actually utilized) and 8GB memory (less than 2.5 GB actually utilized for the largest dataset). In the tested range of number of SNPs, the computation time grew linearly at each sample size (Figure 2). Similarly, for fixed number of SNPs, the computation time grew linearly with the sample sizes. For the largest dataset with 4,000 subjects and 1,000,000 SNPs, it takes about 300 hours for a single process unit. Because the ChromoScan procedure can be naturally divided and paralleled by chromosomes, we reduced the actual wall-clock time for analyzing the dataset to one day using two servers described above (total of 16 processing units). Considering the number of SNPs from available genotyping platforms and the sample sizes of

ongoing GWA studies, a genome-wide scan with 10,000 permutations can be completed in a couple of days with access to multi-core server or computer clusters.

Conclusion

Hundreds of thousands of SNPs were genotyped and tested for genome-wide association studies of common human diseases. Several genomic loci were reported and independently replicated associated with type 2 diabetes [1,2,18–20], Crohn's disease [1,3,21,22] and coronary heart disease [1,23,24], however, the validated loci only explain a small portion of the genetic variance of type 2 diabetes [1,2,18–20] and Crohn's disease [1,3,21,22]. This scan statistic application can detect genomic regions defined by multiple markers considering their distance distribution, and are expected to be robust enough to detect regions without the most significant p-values. In addition, our application is efficiently implemented to allow completion of a genome-wide scan with a large number of permutation tests within hours on a typical computer. We expect to improve our genetic understanding of common diseases beyond single SNP association tests by applying ChromoScan-GWA on more GWA datasets.

Presently, we only calculate one statistic, chi-square, for estimating single SNP association with case-control disease outcomes. We need to incorporate single SNP statistical tests (e.g. linear regression for quantitative traits) in ChromoScan-GWA to scan the whole-genome data effectively. Although our implementation of the permutation test of the scan statistic is very efficient, we hope to improve the whole genome scan performance by replacing the permutation test by a statistic of asymptotic multivariate normal probability which adjusts p-values for multiple correlated tests [25].

A major advantage of the ChromoScan 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. 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 [26,27]. 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 different studies are expected because of differences in the underlying frequency and linkage disequilibrium distribution of SNPs across studies. For GWA studies, the significant SNP associations may not be directly comparable across studies because of different genotyping platforms. Although methods were proposed to impute untyped SNPs across different genotyping platforms [28,29], the imputation errors and increased number of tests (multiple testing) could limit the utility of imputation. Furthermore, for ethnic groups with low levels of LD or that not represented in the HapMap, imputation is not possible without significant introduction of errors. Without introducing additional errors and tests, the ChromoScan can identify the associated regions shared by GWAs using different genotyping platform as long as there are associated SNPs (not necessarily the same SNPs) in the same chromosomal region. In general, the ChromoScan-GWA offers a genome-wide analytical tool for making regional inferences about disease association that is likely to be more robust across studies than single SNP association results.

We developed an analytical strategy for identifying genomic regions associated with drug response by implementing a SNP scan statistic as well as a fast computation application for permutation testing of high-dimensional GWA data. The current ChromoScan-GWA application can be applied for any GWA data with case-control or categorical outcomes to identify associated genomic regions associated. 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. To expand the utility

and the scope of this scan statistic application, we will continue implementing additional single SNP tests, such as regression tests for quantitative traits and interaction tests between SNP and environmental factors. Following these enhancements, we can use the scan statistic for any type of hypothesis test on GWA data – e.g. tests of quantitative traits or tests of gene-environment interaction.

Acknowledgments

This work was supported by National Institute of Health grant HL087660, HL68737, HL 74735 and HL 53335.

References

1. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447:661–678. [PubMed: 17554300]
2. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science* 2007;316:1341–1345. [PubMed: 17463248]
3. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A, Dassopoulos T, Bitton A, Yang H, Targan S, Datta LW, Kistner EO, Schumm LP, Lee AT, Gregersen PK, Barmada MM, Rotter JI, Nicolae DL, Cho JH. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 2006;314:1461–1463. [PubMed: 17068223]
4. Karlin S, Macken C. Assessment of inhomogeneities in an E. coli physical map. *Nucleic Acids Res* 1991;19:4241–4246. [PubMed: 1870978]
5. Wagner A. A computational genomics approach to the identification of gene networks. *Nucleic Acids Res* 1997;25:3594–3604. [PubMed: 9278479]
6. Wagner A. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* 1999;15:776–784. [PubMed: 10705431]
7. Su X, Wallenstein S, Bishop D. Nonoverlapping clusters: Approximate distribution and application to molecular biology. *Biometrics* 2001;57:420–426. [PubMed: 11414565]
8. Hoh J, Ott J. Scan statistics to scan markers for susceptibility genes. *Proc Natl Acad Sci U S A* 2000;97:9615–9617. [PubMed: 10931953]
9. Levin AM, Ghosh D, Cho KR, Kardia SL. A model-based scan statistic for identifying extreme chromosomal regions of gene expression in human tumors. *Bioinformatics* 2005;21:2867–2874. [PubMed: 15814559]
10. Sun YV, Levin AM, Boerwinkle E, Robertson H, Kardia SL. A scan statistic for identifying chromosomal patterns of SNP association. *Genet Epidemiol* 2006;30(7):627–35. [PubMed: 16858698]
11. Sun YV, Jacobsen DM, Kardia SL. ChromoScan: A scan statistic application for identifying chromosomal regions in genomic studies. *Bioinformatics* 2006;22:2945–2947. [PubMed: 17032677]
12. Conover W, Bement TR, Iman RL. On a method for detecting clusters of possible uranium deposits. *Technometrics* 1979;21:277–282.
13. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995;57:289–300.
14. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *J Natl Cancer Inst* 2004;96:434–442. [PubMed: 15026468]
15. Churchill GA, Doerge RW. Empirical threshold values for quantitative trait mapping. *Genetics* 1994;138:963–971. [PubMed: 7851788]

16. CERN. Colt statistical library 1.2.0. <http://dsd.lbl.gov/~hoschek/colt/>
17. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004;74:106–120. [PubMed: 14681826]
18. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson Bostrom K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjogren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Rieke D, Purcell S. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007;316:1331–1336. [PubMed: 17463246]
19. Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S, Baker A, Snorraddottir S, Bjarnason H, Ng MC, Hansen T, Bagger Y, Wilensky RL, Reilly MP, Adeyemo A, Chen Y, Zhou J, Gudnason V, Chen G, Huang H, Lashley K, Doumatey A, So WY, Ma RC, Andersen G, Borch-Johnsen K, Jorgensen T, van Vliet- Ostapchouk JV, Hofker MH, Wijmenga C, Christiansen C, Rader DJ, Rotimi C, Gurney M, Chan JC, Pedersen O, Sigurdsson G, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet* 2007;39:770–775. [PubMed: 17460697]
20. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007;445:881–885. [PubMed: 17293876]
21. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, Green T, Kuballa P, Barmada MM, Datta LW, Shugart YY, Griffiths AM, Targan SR, Ippoliti AF, Bernard EJ, Mei L, Nicolae DL, Regueiro M, Schumm LP, Steinhart AH, Rotter JI, Duerr RH, Cho JH, Daly MJ, Brant SR. Genome-wide association study identifies new susceptibility loci for crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet* 2007;39:596–604. [PubMed: 17435756]
22. Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, Vermeire S, Dewit O, de Vos M, Dixon A, Demarche B, Gut I, Heath S, Foglio M, Liang L, Laukens D, Mni M, Zelenika D, Van Gossum A, Rutgeerts P, Belaiche J, Lathrop M, Georges M. Novel crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet* 2007;3:e58. [PubMed: 17447842]
23. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, Jonasdottir A, Sigurdsson A, Baker A, Palsson A, Masson G, Gudbjartsson DF, Magnusson KP, Andersen K, Levey AI, Backman VM, Matthiasdottir S, Jonsdottir T, Palsson S, Einarsdottir H, Gunnarsdottir S, Gylfason A, Vaccarino V, Hooper WC, Reilly MP, Granger CB, Austin H, Rader DJ, Shah SH, Quyyumi AA, Gulcher JR, Thorgeirsson G, Thorsteinsdottir U, Kong A, Stefansson K. A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* 2007;316:1491–1493. [PubMed: 17478679]
24. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, Hinds DA, Pennacchio LA, Tybjaerg-Hansen A, Folsom AR, Boerwinkle E, Hobbs HH, Cohen JC. A common allele on chromosome 9 associated with coronary heart disease. *Science* 2007;316:1488–1491. [PubMed: 17478681]
25. Conneely KN, Boehnke M. So many correlated tests, so little time! rapid adjustment of P values for multiple correlated tests. *Am J Hum Genet* 2007;81:1158–1168.
26. Patterson M, Cardon L. Replication publication. *PLoS Biol* 2005;3:e327. [PubMed: 16149852]
27. Swarbrick MM, Waldenmaier B, Pennacchio LA, Lind DL, Cavazos MM, Geller F, Merriman R, Ustaszewska A, Malloy M, Scherag A, Hsueh WC, Rief W, Mauvais-Jarvis F, Pullinger CR, Kane JP, Dent R, McPherson R, Kwok PY, Hinney A, Hebebrand J, Vaisse C. Lack of support for the

association between GAD2 polymorphisms and severe human obesity. PLoS Biol 2005;3:e315. [PubMed: 16122350]

28. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 2007;39:906–913. [PubMed: 17572673]
29. Chen WM, Abecasis GR. Family-based association tests for genomewide association scans. Am J Hum Genet 2007;81:913–926. [PubMed: 17924335]

Figure 1.

Region 12 on chromosome 3 associated with the drug response overlaps with SUCLG2 gene. The upper portion shows the arrangement of the genic regions within SUCLG2 and the locations of tagSNPs. The middle portion shows the single SNP p-values of the ten tagSNPs within the identified regions. Significant single SNP tests (chi-square pvalue < 0.05) are shaded in dark color. The lower portion shows pairwise linkage disequilibrium R^2 (numbers in the cells) between the tagSNPs.

Figure 2.

The scalability of the computation speed of ChromoScan-GWA regarding to the sample size and the number of SNPs.

Table 1

Summary of Significant Chromosomal Regions Associated with the Drug Response in Caucasians.

	SNPs w/genotype				Tag SNPs (R ² <0.5)					
	Total SNPs	# of SNP	%	# of SNP p<0.05	% SNP p<0.05	# of SNP	%	# of SNP p<0.05	% SNP p<0.05	# of Regions
Chromosome 1	9200	7590	82.5%	298	3.9%	4485	59.1%	184	4.1%	3
Chromosome 2	10339	8706	84.2%	402	4.6%	4916	56.5%	214	4.4%	6
Chromosome 3	7814	6581	84.2%	297	4.5%	3902	59.3%	173	4.4%	5
Chromosome 4	8563	7049	82.3%	324	4.6%	4014	56.9%	169	4.2%	1
Chromosome 5	8357	7021	84.0%	297	4.2%	4013	57.2%	162	4.0%	2
Chromosome 6	8059	6943	86.2%	301	4.3%	3827	55.1%	156	4.1%	3
Chromosome 7	7041	5949	84.5%	225	3.8%	3361	56.5%	127	3.8%	8
Chromosome 8	6968	5777	82.9%	253	4.4%	3193	55.3%	133	4.2%	4
Chromosome 9	4782	4021	84.1%	169	4.2%	2500	62.2%	99	4.0%	1
Chromosome 10	5673	4769	84.1%	211	4.4%	2823	59.2%	118	4.2%	3
Chromosome 11	5353	4434	82.8%	193	4.4%	2495	56.3%	98	3.9%	4
Chromosome 12	5253	4446	84.6%	210	4.7%	2690	60.5%	124	4.6%	1
Chromosome 13	5214	4325	82.9%	212	4.9%	2470	57.1%	117	4.7%	2
Chromosome 14	4007	3353	83.7%	138	4.1%	1995	59.5%	79	4.0%	0
Chromosome 15	3029	2461	81.2%	102	4.1%	1535	62.4%	68	4.4%	0
Chromosome 16	2378	1941	81.6%	106	5.5%	1332	68.6%	77	5.8%	3
Chromosome 17	1955	1651	84.5%	53	3.2%	1046	63.4%	37	3.5%	0
Chromosome 18	3560	3009	84.5%	130	4.3%	1762	58.6%	65	3.7%	1
Chromosome 19	690	570	82.6%	31	5.4%	418	73.3%	13	3.1%	0
Chromosome 20	2093	1766	84.4%	69	3.9%	1169	66.2%	44	3.8%	1
Chromosome 21	1901	1624	85.4%	57	3.5%	939	57.8%	31	3.3%	0
Chromosome 22	761	623	81.9%	31	5.0%	450	72.2%	22	4.9%	1
Total	112990	94609	83.7%	4109	4.3%	55335	58.5%	2310	4.2%	50

Table 2

Permutation Tests of Significant Regions Identified by ChromoScan-GWA

Region ID	Chr.	Region Start	Region Size	tagSNP#	Significant tagSNP#	Scan Statistic p-value	Empirical p-value
1	1	61546517	36200	8	3	2.983E-03	0.55747
2	1	81100856	79607	4	3	4.819E-03	0.70050
3	1	215332585	1032	3	2	5.527E-03	0.74370
4	2	34196503	257878	10	4	4.901E-03	0.48423
5	2	42414681	22278	3	3	7.008E-04	0.13262
6	2	53908234	415956	9	4	9.281E-03	0.69445
7	2	123864287	227570	8	4	2.103E-03	0.27517
8	2	124731189	485	3	2	3.305E-03	0.37475
9	2	159261446	36725	3	3	1.809E-03	0.25058
10	3	1144496	2364	3	2	9.012E-03	0.65128
11	3	6934982	262254	11	4	5.479E-03	0.49458
12	3	67392670	142911	10	5	6.295E-05	0.02880*
13	3	141131024	1916	3	2	8.517E-03	0.63100
14	3	147785484	67336	7	3	6.218E-03	0.53654
15	4	189670188	7608	4	3	2.183E-04	0.23920
16	5	96125021	356892	10	5	1.274E-03	0.41610
17	5	154758389	160838	7	4	2.075E-03	0.53765
18	6	66434595	268672	8	4	4.532E-03	0.62400
19	6	85169956	90502	5	5	1.231E-05	0.01637*
20	6	137760226	39061	5	3	2.847E-03	0.48030
21	7	7123923	50224	4	3	3.399E-03	0.20509
22	7	20202393	84216	6	3	7.748E-03	0.34930
23	7	21682204	84696	4	3	6.378E-03	0.30756
24	7	35278444	501160	11	5	2.058E-03	0.14705
25	7	67494770	889	3	2	5.532E-03	0.27838
26	7	79375212	49731	5	3	3.876E-03	0.22153
27	7	92271648	26337	6	3	1.728E-03	0.12951

Region ID	Chr.	Region Start	Region Size	tagSNP#	Significant tagSNP#	Scan Statistic p- value	Empirical p- value
28	7	135895633	252125	10	5	4.935E-04	0.06015
29	8	26511632	189771	10	4	3.118E-03	0.35245
30	8	70886578	1263	3	2	7.057E-03	0.57658
31	8	78573650	282349	7	4	5.675E-03	0.51405
32	8	140486231	138097	10	4	1.552E-03	0.22545
33	9	89978360	400885	7	4	9.669E-03	0.90550
34	10	4501571	189397	9	4	2.932E-03	0.35663
35	10	61821021	313458	10	5	6.048E-04	0.12997
36	10	90012627	25502	5	3	1.357E-03	0.20867
37	11	19960688	1340	3	2	6.172E-03	0.34218
38	11	82867048	219562	6	4	2.305E-03	0.18924
39	11	96171790	829	3	2	4.817E-03	0.29724
40	11	98793413	2962	3	2	9.633E-03	0.45036
41	11	106424364	223983	7	4	2.270E-03	0.18730
42	12	21063320	542553	11	6	4.485E-04	0.28470
43	13	72876727	151081	10	6	9.965E-06	0.01045*
44	13	107520582	30972	5	3	2.678E-03	0.40420
45	16	12509416	112479	11	4	8.984E-04	0.11423
46	16	63452171	87363	5	3	7.259E-03	0.40647
47	16	81379173	341	3	3	3.291E-06	0.00507*
48	18	3317801	20012	4	3	7.355E-04	0.20720
49	20	37551920	336796	10	5	6.803E-04	0.14360
50	22	32200000	58933	4	3	2.095E-03	0.10730

* Significant Empirical p-values less than 0.05.