# WHOLE GENOME ANALYSIS OF SINGLE NUCLEOTIDE POLYMORPHISM ALLELE FREQUENCY AND FALSE POSITIVE RATE

by

**Pei-Chien Tsai**

B. S. in Public Health, Kaohsiung Medical University, Taiwan, 2003

M. S. in Public Health, Kaohsiung Medical University, Taiwan, 2005

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Pei-Chien Tsai

It was defended on

July 31, 2009

and approved by

Eleanor Feingold, Ph.D, Associate Professor
Department of Human Genetics, Graduate School of Public Health
University of Pittsburgh

Chien-Cheng (George) Tseng, ScD, Associate Professor
Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

**Thesis Advisor**: M. Michael Barmada, Ph.D, Associate Professor
Department of Human Genetics, Graduate School of Public Health
University of Pittsburgh

# WHOLE GENOME ANALYSIS OF SINGLE NUCLEOTIDE POLYMORPHISM ALLELE FREQUENCY AND FALSE POSITIVE RATE

**Pei-Chien Tsai, M. S.**

**University of Pittsburgh, 2009**

Genome-wide association (GWA) studies are used widely for detecting gene variants' contribution to diseases and traits. Recent researches indicate to several methodological challenges in the study design for GWA, for example, sample size issues, power calculations, false positive rate adjustments, and commercial chips' coverage of the genome. Chromosomal regions can also influence the observed genetic diversity under certain conditions; mainly the regions of secondary structures and large-scale repeats may affect the fidelity in marker genotyping. This study was to find such regions that contained markers with more variability and to examine the correlation of this variability to the factors relevant in a GWA study design, such as the false positive rate. We enrolled healthy controls from eight independent GWA designs then assigned randomly into case and control status. Minor allele frequency estimates, and case-control association analyses were performed using PLINK for sets with different sample sizes. Marker numbers exhibiting high variability in the allele frequency estimates, and the average number of false positives were calculated for bins across the autosomal genome. We found that SNP variability (in allele frequency) was unrelated to the sample size. More variable regions correlated to regions of more average number of false positives, after adjusting for confounders, such as sample size. We suggested that regions with more variability might have structural characteristics that made them difficult to be scanned during the genotyping process. Our study has great public health relevance because regions with more variability could undermine the

effective study of a candidate genes and disease relationship during a research, or worse leading to erroneous conclusions. We advise in studying these regions, the researchers could lower their false positive rates to avoid inaccurately significant levels.

## TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ACKNOWLEDGMENT

I confer my highest gratitude to my advisor, Dr. M. Michael Barmada, who went through much effort building my knowledge for this study, as well as for me to learn the required computer program from a preliminary level, and his great advice in solving problems. This study could not be accomplished without his full help, support and patience. I also like to thank my academic advisor, Dr. Eleanor Feingold, who helped with all matters academic and gave great suggestions. I also like to thank Dr. Barmada, Dr. Feingold and Dr. Chien-Cheng (Geroge) Tseng for being my committee members during my thesis defense, and gave many valuable opinions to this study. During my Masters study, I had the fortune to participate in lectures by brilliant professors from Biostatistics and Human Genetics department, not only granted me understanding, also inspired me for my future life.

Thank you so much Ya-Hsiu (Ami) Chuang, Chia-Ling (Carry) Kuo, and Shuya Lu for your programming support and always willing to help. My classmates and friends, especially Akunna who corrected English in my thesis and Tzu-Ching, thank you all it has been great to have you around. Finally, I would like to give my appreciation to Albert, for his whole-heartedly support in every aspect of my life and study, especially for his great patience in helping me out everything. My parents, brother, aunt, and other family members have been my strong personal support all the time, thank you for making me brave overseas and hold my back. I dedicate this thesis especially to my mother and father, and my God, any little accomplishment by me shall be your glories all the time.

# 1.0 INTRODUCTION

## 1.1 HISTORICAL REVIEW OF MOLECULAR GENETICS

In early nineteenth century, scientists questioned whether the phenotype of an individual was related to disease status. The first twin study in the world, Jablonski's[1, 2] study on eye among 52 twins showed identical twins with fewer phenotypic differences than non-identical twins. Hermann Siemens[3] suggested hereditary factors could be assessed in identical and non-identical twins from his study on skin naevi. Merrriman's[4] publication of intelligence quotient (IQ) exam using psychological monographs on a twin study, found identical twins had higher correlations in IQ. Idea was that identical twins had more resemblance since sharing origins from egg and arrangement. These studies outlined a vague concept of physical genetics. True source of person to person variation remained unknown.

In 1928, Frederic Griffith found genetic information transferrable from dead to living bacteria and activation in new host. Later in 1944, Oswald Avery, Colin McLeod, and Maclyn McCarty discovered DNA's responsibility for this transfer in the bacteria[5]. Alfred Hershey and Martha Chase proved this theory in 1952 that DNA carried the genetic information for inheritance[6]. James Watson and Francis Crick discovered the double helix DNA structure on February 28, 1953[7]. They showed genetic information in the nucleotides of both strands and inheritance by duplication of each DNA strand. A decade apart in 1975, Edwin Mellor Southern developed the southern blot method[8]. By this method, a specific sequence within DNA could be presented using molecular hybridization. Same year, Allan Maxam and Walter Gilbert developed DNA sequencing using radiolabeling[9]. The chain-terminator method later replaced the sequencing method, developed by Frederick Sanger in 1975[10], where dideoxynucleotide triphosphates were used as DNA chain terminators.

A breakthrough invention was Kary Muliis who developed the polymerase chain reaction (PCR) in 1983 based on the original thought from H. Gobind Khorana and Kjell Kleppe in 1971. Using this method, DNA sequence could be amplified into large amounts and product stored stably for an extended time for later treatments, such as restriction fragment length polymorphisms (RFLP), developed by Alec Jeffreys in 1984. PCR was later modified for different genetic studies. For example, Real-time PCR, Q-PCR, and reverse transcription PCR (RT-PCR). Three years later, Leroy Hood pioneered four instruments: DNA gene sequencer and synthesizer, and protein synthesizer and sequencer, making the study of genetics faster and easier.

In 1990, the United States Congress approved of fifteen years human genome project (HGP), originally scheduled for completion by end of 2005. By 1995, John Craig Venter finished whole genome sequencing in bacteria. Surprisingly, the human genome project announced in 2003 that preliminary sequencing was done for each chromosome. Three years later, the gaps were filled and details had been provided for each chromosome. About same time, the whole genome sequencing was completed for the mouse in 2003 and rats in 2004.

## 1.2 VARIATION AMONG HUMAN POPULATIONS

There are two sources of physical variation: environmental influence and gene variation. Gene variation can be contributed by mutation during gene DNA replication, and recombination during sexual mating, such as crossing-over switch in gene loci. Improvements to molecular genetic techniques in late nineteenth century raised naturally questions as how genes related to disease and regulated disease processes. Several types of gene variation exist. Two primary types are single nucleotide polymorphism and copy number variation. Both can affect gene activity to further express traits or disease.

The single nucleotide polymorphism (or SNP) is a single nucleotide variant of DNA sequence differing between specie members. SNP sites contain typically two alleles, where the less frequent is a variant or mutant or minor allele. SNPs are located on any regions of a chromosome, in the coding or non-coding or intergenic regions. If located in a coding sequence, they can change amino acid sequence (known as nonsynonymous polymorphism) with correspondent change to protein function to be missense (replacing one amino-acid with another) or nonsense (producing an aberrant stop codon). If they do not change the amino acid sequence, it is a synonymous or silent variant. SNPs have MAF (minor allele frequencies) that vary with ethnicity (known as population SNPs; admixture SNPs; ethnicity SNPs), gender (specifically SNPs on the sex chromosomes and others[11]), and disease status. For these reasons, SNPs are widely used for disease association studies. There is good evidence that tagging SNPs (tSNPs, proxy markers for neighboring genetic variation) and combined effect of several SNPs (i.e. a haplotype) can increase disease susceptibility or disease protection. Increasingly, this can be used to personalize medicine, increasing treatment efficacy and minimize adverse drug reactions.

The copy number variation (CNV) is a number of nucleobase copy segments of DNA sequence with sizes varying from kilobases to megabases. Meiosis can cause CNVs, where DNA undergoes recombination and genomic rearrangement. Deletion, duplication, inversion, and translocation of nucleobases may occur. Kidd et al[12] has mapped eight human genomes to find structural variation. The mapped 6.1 million clones have nearly 0.4% genome difference from person to person, and there are regions with more CNVs such as regions with deletions, insertions, and inversions within each chromosome.

Previous studies have shown SNPs and CNVs associate with human phenotypic variability, disease-related trait, and disease susceptibility. Thus studying single nucleotide or CNV may find the disease-related genes based on accepted or suspected pathomechanisms, however this gene selection method can be biased.

## 1.3 GENOME-WIDE ASSOCIATION STUDY

Genome-wide association study (GWAS) represents an unbiased examination of possible contributions of gene variants to a specific disease or a trait. After genotyping or sequencing of participants, SNP frequencies are gathered and compared across the case and control groups. If the minor allele frequency (MAF) of a SNP deviates significantly between cases and controls using an appropriate statistical test, the SNP and its region are considered to associate with the disease or trait. Despite advantages in GWAS design (ability to find novel gene variants and disease relationships, thorough investigation of variation across genome, no prior information required of gene function or family information), it is often impractical due to the extremely high cost for expanded sample sizes and detection of rare allele (MAFs <1%).

## 1.4 METHODOLOGICAL CHALLENGES IN GENOME-WIDE ASSOCIATION STUDY

Genome-wide association studies since development in 2002 are not without methodological problems that need to be addressed. Three major problems are the study design of GWAS, the replication issue, and the gene-gene interaction. One essential point in study design is the power issue. Power in GWAS design is influenced by: false positive rate (type I error, α), type II error (β), relative risk of case and control group, and sample size. Jorgenson and Witte has indicated to achieve power of 80% (usually set as the criterion for Epidemiology research), the required sample size is approximately 1800-2000, α level= $10^{-5}$-$10^{-6}$, odds ratio more than 1.5, and disease-allele frequency ranging from 0.3-0.7 (in present study should be 0.3-0.5 for minor allele frequency)[13].

The definition of type I error is the error rate to reject the null hypothesis given it is true, whereas type II error (β error) is the error rate to reject alternative hypothesis given it is true. β error in genome-wide association study can be represented as β (n,m,y), where n is the sample size; m is the maximum $r^2$ value; and y for other parameters affecting power (effect size, disease-allele frequency, etc). In a genome-wide association study, type I error is adjusted to a lower level, by dividing it by the total number of independent tests[14] (known as the Bonferroni post hoc correction). Unadjusted type I error will increase the false positive rate, resulting in associations between SNPs and disease or quantitative traits to be caused by chance. For genome-wide association studies, comparisons are usually large enough to affect the significance level ranging from 0.05 to $10^{-4}$-$10^{-7}$.

Furthermore, despite Bonferroni correction being used widely for type I error adjustment, it can overestimate the significance level if SNPs are highly correlated with each other (such as locations of high linkage disequilibrium (LD)). Then number of tests will be less than the number that is supposed to be assayed[15]. To adjust for the possibly highly-related SNPs (those in linkage disequilibrium), tag SNPs are selected as the one SNP to represent the whole related cluster of SNPs in a small region[16]. Linkage disequilibrium between SNPs is used to find tag SNPs. Haplotype frequency between two SNPs is calculated using expectation maximization method, where $r^2$ is subsequently computed as the correlation of alleles at two sites. SNPs with highest $r^2$ will be taken as the tag SNP. The meaning of $r^2=1$ means one marker can cover complete information over the other marker, which means perfect LD.

Other issues, such as the commercial chips' coverage, study population, disease allele frequency, imputation of missing value, and analysis methods may influence the study power. Based on ways of SNP selection (e.g. scan distance and scan numbers), the commercial chips' coverage can be different. Some chips seem to be more effective in specific population. For example, Illumina chips are optimized on CEU HapMap data (Utah residents with ancestry from Northern and Western Europe). To date, all commercial chips tend to lose some power in African-American populations. Thus African-American populations study requires more sample size, or higher relative risk, to achieve higher power. Populations with less linkage disequilibrium or less racial integration may have allele frequency different to other populations[17].

SNPs of disease allele frequencies less than 1% are suggested for removal prior association analysis. Disease allele can have either risk or protection effect. Sometimes one SNP site has an opposite effect on a disease due to different study populations. Recommended rare minor allele frequency ranges from 0.3-0.5. Most commercial chips are not effective for SNPs with lower minor allele frequency. Missing genotypes influence the result of an association study. SNP imputation can be based on the outer database, such as HapMap database. The gene frequencies can be imputed using outer database without losing too much power[18], as Hao et al's research have shown by marking out 1% known SNPs as 'untyped' SNPs, imputed the SNPs using outer database then compared the imputation SNPs to observed SNPs.

There is a problem that researchers can hardly replicate results from GWAS design even when studying the same disease. Any single variable, such as the choice of commercial chips, study population, control of false positive rate and β error, and so on, can influence results. Another issue is we may detect the SNP sites that associated with the diseases or traits, but the gene-gene interaction effect or the gene-environment interaction effect are undetectable using GWAS design due to a large number of SNPs' comparisons.

## 1.5 STUDY PURPOSE

Present study was performed to study the false positive rate in the control populations in a GWAS design. This study explored two major interests: 1) were there any locations in a genome that exhibited an unusual skewing of allele frequency using random sample sets; 2) how do those regions of SNP site with higher standard deviation numbers influenced the false positive rate in the same region?

# 2.0 MATERIALS AND METHODS

## 2.1 DATASETS

Subjects enrolled into this study were healthy controls and isolated subjects obtained from the following eight independent GWAS designs.

***International Multi-Center ADHD Genetics project of GAIN, from dbGaP[19]***

This study focuses on the associated SNP markers with Attention Deficit Hyperactivity disorder (ADHD), analysis of quantitative ADHD phenotypes, complete copy number analyses, assessment of parent of origin and season of birth effects. Also, testing for epistasis among uncorrelated genes. A total 600,000 tag SNPs have been scanned in this international multisite ADHD Genetics (IMAGE) project (ADHD2). Subjects have been recruited from twelve centers over eight countries: Belgium, Germany, Holland, Ireland, Israel, Spain, Switzerland, and United Kingdom.

***Major Depression: Stage 1 Genome-wide Association in Population-based sample, project of GAIN, from dbGaP[19]***

This study is to identify the genomic regions that help protect or confer susceptibility from MDD (major depressive disorder). This study has involved 1780 cases and 1860 controls.

Cases are drawn from the Netherlands Study of Depression and Anxiety (NESDA)[20] and controls from the Netherlands Twin Registry (NTR)[21].

### *Search for Susceptibility Genes for Diabetic Nephropathy in Type 1 Diabetes (GoKinD study participants), project of GAIN, from dbGap[19]*

This study intends to identify the susceptibility genes for diabetic nephropathy in type 1 diabetes. All subjects (n= 3043) are diabetic patients, subdivided into two sets: those with kidney disease and those with normal renal status in spite of long-term diabetes. Source data has also been from the NIDDK (dbGaP phs000088 Search for susceptibility Genes for Diabetic Nephropathy in Type 1 Diabetes).

### *POPRES: Population Reference Sample, project of NHGRI, from dbGaP[19]*

Ten unrelated populations of nearly 6,000 subjects of African-American (USA), East Asian (Taiwanese and Japanese), South Asian (Indians), Mexican (Mexico), and European origin (Australia, Canada, United Kingdom, Switzerland, USA) are included in this study and are all healthy controls. Genotyping has been performed by Affymetrix GeneChip 500K array set and 96-well-plate format[22].

### *NINDS Parkinson's disease, project of NINDS, from dbGaP[19]*

This study is to identify those gene variants exerting a large effect in risk for Parkinson's disease in a population cohort. Additionally, generate a public genome-wide genotype data. There are 270 neurologically normal controls, and 408803 unique SNPs have been detected using the Illumina Infinium I (109365 gene-centric SNPs) and Infinium I HumanHap 300 assays (317511 haplotype tagging SNPs)[23].

### *HumanHap550 kv3/kv1, from Illumina iControl DB*

Illumina I ControlDB is a collection of online databases that provides control subjects' information of genotype and phenotype. It is freely accessible by Illumina users. Of the six online databases, only HumanHap550 has been included in our study. There are 5566 subjects in this database, representing 51 populations worldwide, majority are Caucasians (n= 3304) and African-American (n= 1902). Others are Asian, Hispanic, African, and American Indian. Two, Illumina HumanHap 550-Duo BeadChip kv3 and kv1, have detected SNPs and enabled to detect more than 555,000 tag SNP markers (data were downloaded from http://www.illumina.com/pages.ilmn?ID=231).

### *HapMap 270, from International HapMap Project*[24]

This project focuses on common patterns of genetic variation in humans that includes SNPs, tSNPs, and Haplotypes. From different populations 270 subjects are included, such as Nigeria (Yoruba, n= 30, both parent and adult-child trios), Japan (n= 45, unrelated individuals), China (n= 45, unrelated individuals) and U.S. (n= 30, residents with ancestry from Northern and Western Europe by Centre d'Etude du Polymorphisme Humain (CEPH)). Samples have been genotyped for at least one million SNPs across the human genome. Kits from each of centers detect the SNPs, such as Third Wave, Ilumina, PerkinElmer, Sequenom and ParAllele.

***Human Genome Diversity Project***[25-28]

This project aims to find diversity and unity in the human race. Main components are data/sample collection, preservation, analysis, database creation and management. Samples are cell-lines or DNA samples obtained from populations, maintained at the Centre Etude Polymorphism Humain in Paris, France. In defining the human populations, language has been used as the primary criterion. The project aims to collect 500 out of an estimated total of 4000 to 8000 distinct populations within five years.

## 2.2 DATA MANAGEMENT

We selected eligible subjects from the eight original datasets aforementioned. Transposed pedigree files (.ped) and transposed family files (.fam) were created when they were required for PLINK analysis. All datasets were merged using PLINK[29]. Prior to merging, flip alleles were checked in each database and flipped properly. Then PLINK created .bed files, .bim file, and .fam file.

Appendix and Figure 2 show the data management procedure. To perform the association study, we assigned subjects randomly into cases and controls, and selected randomly for future analysis. Both phenotype list and subject lists from the merged datasets were generated using R. Subjects were selected randomly to form different sample sizes of 25, 50, 75, 100, 150 and 200, then replicated for 500 times. The new selected subject ID list was kept in PLINK while running the frequency estimates and the phenotype list would be added in association study, which concluded with 500 frequency and association files.

Minor allele frequency columns and minor allele number in each frequency files were created separately into two files. Allele number file was used for criteria selection, because not all subjects in our datasets contained full information of SNP frequencies due to different commercial chips used in each study. Although more replicates in one SNP site could give more stable and reasonable standard deviation, it could also lose much information. We decided to keep those replicates with more than half non-missing allele number. Those replicates with minor allele number less than 50% of two times the total selection number was set as 'NA' in the allele frequency file. Since the regions of interest were with unusual higher standard deviation or higher average number of false positives than the effect of single SNP, all SNPs were grouped as 'bins' using their SNP location (bp), bin number will be assigned for every 10,000,000 bp. The longest chromosomes were chromosome 1 and chromosome 2, which contained 25 bins (~25,000,000 bp), and the shortest chromosomes were chromosome 21 and chromosome 22, which contained only 5 bins (~6,000,000 bp). After a preliminary analysis of our dataset, chromosome 23 contained noisy information compared to other chromosomes, thus we decided against discussion of the variability of chromosome 23 in our study.

Our data analysis consisted two major parts: regions of higher standard deviation number detection and average number of false positives detection.

### 2.2.1 Regions of higher standard deviation detection

Regions that contained more numbers of SNP sites with higher standard deviation drew our attention. Observations were made as how it distributed in chromosomes and varied with sample sizes. There were two steps for picking up higher standard deviation numbers. Firstly, picked up SNP sites that had higher standard deviation of their 500 replicates were compared to other SNP sites. Secondly, SNPs with higher standard deviation (SD) was counted for each bin in each chromosome. Our definition of higher standard deviation was:

$$\begin{aligned}&SD\ of\ 500\ replicates\ in\ one\ SNP\ site >\\&average\ SD\ of\ the\ chromosome + 2 \times SD\ of\ the\ chromosome's\ SD\end{aligned}$$

For example, a SNP site *X* in chromosome 1 would be counted as a higher SD location if *X*'s SD > average SD of chromosome 1 + 2 × SD (the SD of each SNP on chromosome 1). Those SNP sites with higher SD were plotted with their SNP location (bp) to see the specific skew or cluster in any region. The comparison of SD in every two chromosome was performed using t-test, and the equal variance test was performed prior to t-test. Box-plot of SD was plotted for all chromosomes.

### 2.2.2 Average number of false positives detection

Each subject underwent random labeling as cases or controls for the association analysis using PLINK. Since all subjects were originally healthy controls, no difference in their minor allele frequencies and non-significant p-value were expected. The emphasis was on the regions that contained unusual higher average number of false positives. The selected samples were the same samples used in the previous 'regions of higher standard deviation' therefore, the association analysis also ended up with 500 replicates. P values of each association analysis were considered differently as being 'significant' depending on the sample size. For sample size 25, 50, 75, 100, 150 and 200, the criteria for significance were 0.05, 0.01, 0.005, 0.001, 0.0005, and 0.0001 respectively. Replicates with significant levels were counted for each SNP site as the ‘number of significant replicates’, and the average ‘number of significant replicates’ was counted for each bin in each chromosome as the ‘average number of false positives’.

The resulting dataset contained information on the sample size, chromosome number, bin number, higher SD numbers, and average number of false positives. Pearson’s correlation test was then performed to detect the correlation between those factors to see the relationship between each factor. To find the factors that could better predict the average number of false positives, univariate and multivariate regression was used. Univariate regression was performed for each factors and average number of false positives first, those factors with statistical significance in the univariate regression model would be placed into the multivariate regression model.

## 2.3 STATISTICAL ANALYSES

The statistical programs used in our study, R, Unix system, and PLINK were executed on the online server GATTACA (Human Genetics Grid Cluster). GATTACA is a computing grid in the Human Genetics department, University of Pittsburgh Graduate School of Public Health that allows its users to work the gene study online. R was used in our data-handling, preliminary statistics, and plotting. PLINK[29] is a whole genome association analysis toolset that is applied for genotype or phenotype data. PLINK can be used for data management, basic association testing, and multimarker predictors. In this study, PLINK was used for database merging, subsets extracting, strand flipping, frequency estimation and case-control association study. Two sample t-test was performed for comparing the mean of SD in the pair-wise chromosomes, box-plots were performed for the spread of SD in each chromosome, dot plots were performed for cluster or skewing detection in each chromosome, using R. Correlation between each factor was tested using Pearson Correlation analysis. Univariate and Multivariate was performed for predictors' detection using R. In present study, P value was significant depending on different criteria. For average number of false positives detection, the criteria of P value depended on the sample size. In the pair-wise t-test analysis, we did multiple comparison 231 times ($\binom{22}{2} = 231$) and increased the type I error, thus Bonferroni's correction was performed to lower the $\alpha$-level from 0.05 to 0.00216.

# 3.0 RESULTS

## 3.1 SNP SITES WITH HIGHER STANDARD DEVIATION OF MINOR ALLELE FREQUENCY

The distributions of standard deviation of minor allele frequency in each chromosome for different sample sizes are shown on Figures 3-1(a) to 3-6(a). There were more outliers in sample sizes 25, 150, and 175 and fewer outliers in sample sizes 50, 75, and 100. Despite the mean of standard deviation appearing similar or only slightly different in decimal (Figures 3-1(b) to 3-6(b)), the pair-wise t-test results indicated nearly half of chromosomes differed from each other (Figures 3-1(c) to 3-6(c)). Overall, trends in average standard deviation of minor allele frequency for each chromosome were similar using different sample sizes (Figure 1).

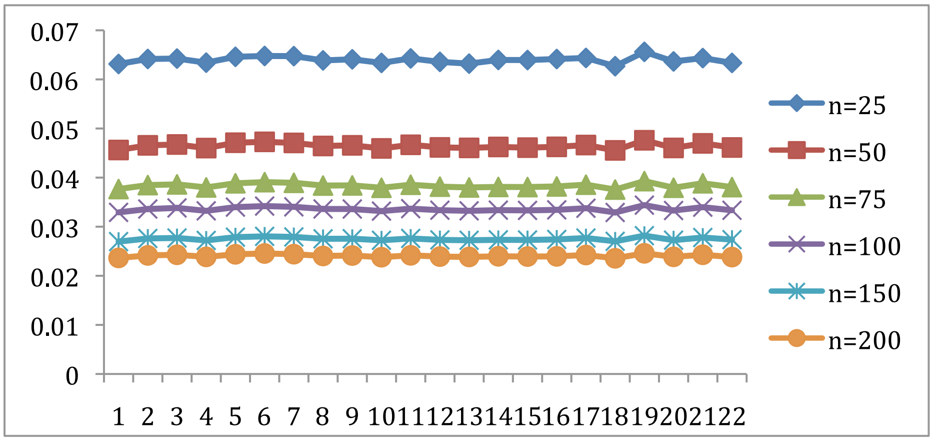


**Figure 1. Trends in average standard deviation of minor allele frequency for each chromosomes using different sample size**

The outlier of standard deviation of minor allele frequency on same chromosome varied with different sample size. For example, in sample sizes 25 and 50, there were extremely higher outliers located on chromosome 14 and 16, but similar results did not show for other sample sizes. In sample size 50, there were extremely higher outliers on chromosome 3, 5, 14 and 16. In sample size 75, there were extremely higher outliers on chromosome 4, 5, 6, 7, 14, 15, and 22. In sample size 100, there were extremely higher outliers on chromosome 1, 3, 9, 11, and 20. In sample size 150, there were extremely higher outliers on chromosome 3 and 8. In sample size 200, there were extremely higher outliers on chromosome 2. Although some chromosomes like chromosome 5, 14, and 16, tended to have higher variability in standard deviation, results were not consistent using different sample sizes.

Figures 4-1 to 4-22 show the dot plots of outliers and locations on chromosome 1 to 22 using different sample sizes. The upper left to lower right represented the outliers for sample size 25, 50, 75, 100, 150, and 200. We could observe some outliers might not have the same extent each time with different sample sizes, but those SNPs still seemed to have higher standard deviations all the time.

## 3.2 NUMBER OF FALSE POSITIVES DETECTION

Replicates with P values higher than sample size-dependent significance level were counted for each SNP site as the ‘number of significant replicates’. The average numbers of significant replicates were counted for each bin as the ‘average number of false positives’. We had a dataset that contained information on sample size, chromosome number, bin number, higher SD numbers, and average number of false positives. The average number of false positives was interpreted as the outcome, other factors considered as possible predictors.

Scatter plots of each predictor and average number of false positives are shown in Figures 5-1 to 5-4. As sample size increased, the average number of false positives decreased due to adjusted $\alpha$ level (Figure 5-1). The average number of false positives did not change much in different chromosome or bins (Figures 5-2 and 5-3), but seemed to change with the number of higher standard deviation (Figure 5-4). Pearson's correlation test was then performed to observe the correlations between each factor (Table 2). The average numbers of false positives highly correlated to the sample size (r= -0.687) and the number of higher standard deviation (r= 0.752).

For univariate regression analysis (Table 1), sample size and number of higher standard deviation were significant predictors to the average number of false positives. Both of them had high coefficient of determination, $R^2$, to explain outcome. Chromosome and bin effect were excluded from the final multivariate model for insignificance in the univariate model. In the final multivariate regression model, the overall P value for this model was significant. With the sample size and number of higher standard deviation effects kept in the model, this gave us the regression model:

$$Y(\textit{average number of false positives}) =$$
$$8.676 - 0.069 \times X1(\textit{sample size}) + 0.609 \times X2(\textit{number of higher SD})$$

# 4.0 DISCUSSION

Few SNP markers had more variability in their minor allele frequencies on chromosomes. As we selected randomly all healthy controls from a large dataset, and selection repeated many times (500 replicates), the standard deviation of these replicates were expected to fall into a reasonable range. However, few still had more variability in their allele frequencies, and this variability might translate into unreliable results. So we examined whether this variability was sufficiently large to create a region, where false-positive signals were more common compared to other regions. Equal sample size of case-control association analysis was performed to find unusual markers that had higher number of false positives, and regions containing these markers compared to those regions with more variability.

The regions with more variability correlated highly to the regions where false-positive signals were more common, after adjusted for other confounders such as sample size. In our final multivariate regression model, we could see that average number of false positives increased with number of higher standard deviation, but decreased with sample size. As the P value depended on sample size, it was reasonable to get the negative relationship between sample size and outcome. Compared to the sample size effect, the number of higher standard deviation had more effect on outcome. This result indicated that more variable regions could influence the false positive rate in the same region.

Rapid progress in genome-wide association studies has caused researchers to focus on power issue or sample size while neglected the possibility that some markers could influence the study result for its variability. Based on previous study, commercial chips may be advantaged in saving time and making the experiments easier but may have coverage problem or other unknowns. In our study, we used different sample size of subjects and ran a large number of replicates to make sure our results were not given by chance, but still found some regions had a wide range of minor allele frequency variability within it, and suggested these regions were difficult to be scanned. Technical difficulties could happen in genotyping process, for some regions could have large copy-number variations or the DNA structure was difficult to be amplified. Both of these situations can contribute to human genetic diversity[30] and cause errors while doing the SNP detection.

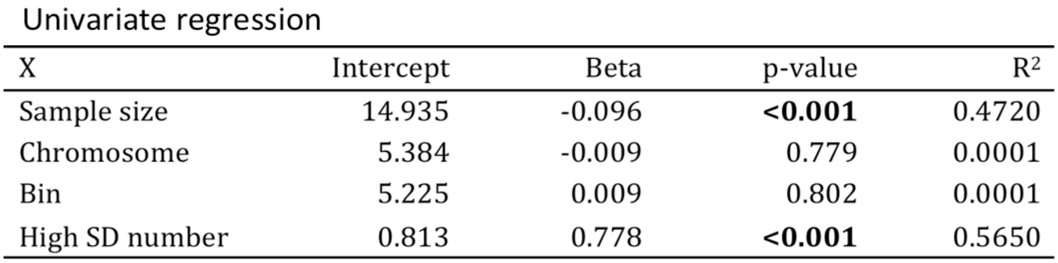
We proved that markers in some regions did have more variability, and those regions could easily associate with disease in an association study if the false positive rate in that region had not been adjusted. Based on our speculation, more variable regions might be with second structure problems or large-scale repeats, CNVs. We compared our outliers' plots to Itsara[30]'s recent result chromosome by chromosome (Figures 6-1 to 6-7). Though the regions with more CNVs in their studies might not perfectly match the regions with more variability in our study, we still observed the similarities between the more variable regions in our plots and their 'predicted rearrangement hotspots' that were considered to be associated with diseases in their study.

# 5.0 CONCLUSION

We found more variable regions on chromosomes corresponded significantly to increases in the average number of false positives. This study supports the understanding that there is possibly a structural source of error from genome-wide association techniques. Our result raised another important problem for consideration besides sample size and power. We advise in studying the genetic regions with more variability, the researchers could lower their false positive rates to avoid inaccurately significant levels. The appropriate amount to be confirmed is an area for further research.

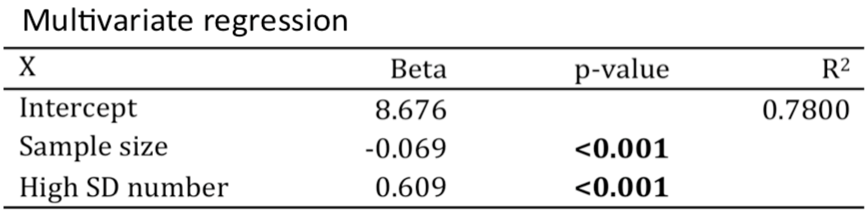
# 6.0 TABLES

**Table 1. Univariate and multivarite regression of average number of false positives and predictors**



Univariate regression

| X | Intercept | Beta | p-value | $R^2$ |
|---|---|---|---|---|
| Sample size | 14.935 | -0.096 | **<0.001** | 0.4720 |
| Chromosome | 5.384 | -0.009 | 0.779 | 0.0001 |
| Bin | 5.225 | 0.009 | 0.802 | 0.0001 |
| High SD number | 0.813 | 0.778 | **<0.001** | 0.5650 |



Multivariate regression

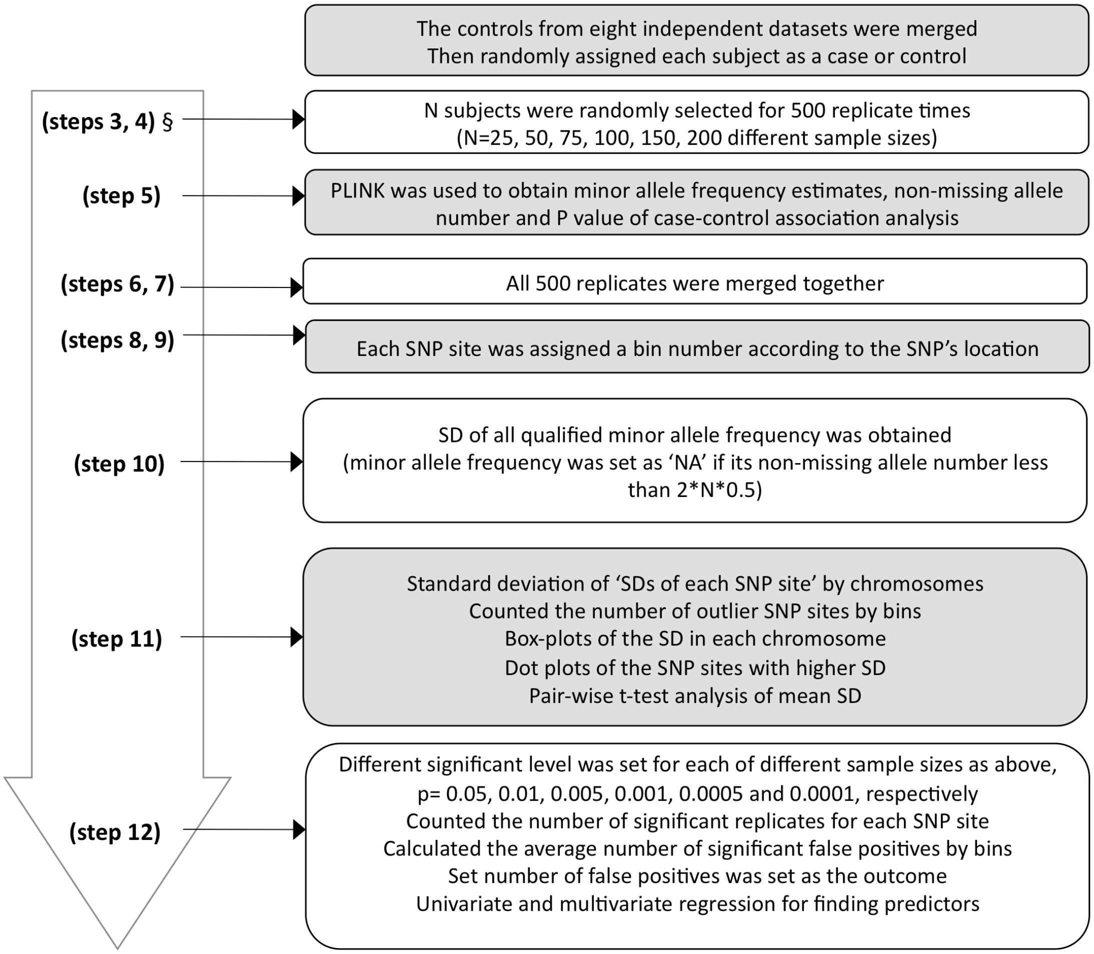
| X | Beta | p-value | $R^2$ |
|---|---|---|---|
| Intercept | 8.676 | | 0.7800 |
| Sample size | -0.069 | **<0.001** | |
| High SD number | 0.609 | **<0.001** | |

**Table 2. Pearson's correlation coefficient between each factor**

| (N=1764) | Sample size | Chromosome | Bin | High SD | NFP |
|---|---|---|---|---|---|
| Sample size | | | | | |
| r[a] | 1 | 0.000 | 0.000 | **-0.333** | **-0.687** |
| p-value[b] | | 1 | 1 | < **0.001** | < **0.001** |
| Chromosome | | | | | |
| r[a] | 0.000 | 1 | **-0.485** | 0.003 | -0.007 |
| p-value[b] | 1 | | < **0.001** | 0.915 | 0.779 |
| Bin | | | | | |
| r[a] | 0.000 | **-0.485** | 1 | -0.046 | 0.006 |
| p-value[b] | 1 | < **0.001** | | 0.052 | 0.802 |
| High SD | | | | | |
| r[a] | **-0.333** | 0.003 | -0.046 | 1 | **0.752** |
| p-value[b] | < **0.001** | 0.915 | 0.052 | | < **0.001** |
| NFP | | | | | |
| r[a] | **-0.687** | -0.007 | 0.006 | **0.752** | 1 |
| p-value[b] | < **0.001** | 0.779 | 0.802 | < **0.001** | |

[a] r: Pearson correlation coefficient; [b] p-value: Pearson correlation significance (2-tailed); High SD: number of high standard deviation ; NFP: average number of false positives

# 7.0 FIGURES

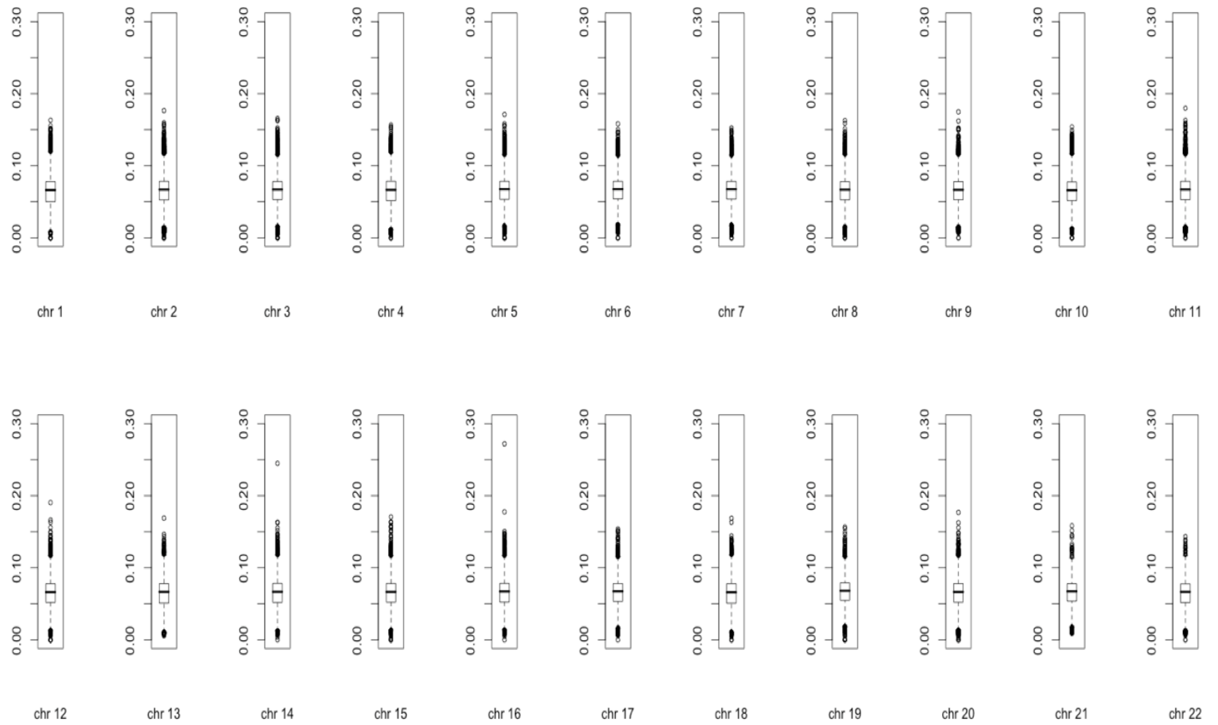


The controls from eight independent datasets were merged
Then randomly assigned each subject as a case or control
(steps 3, 4) §
N subjects were randomly selected for 500 replicate times
(N=25, 50, 75, 100, 150, 200 different sample sizes)
(step 5)
PLINK was used to obtain minor allele frequency estimates, non-missing allele number and P value of case-control association analysis
(steps 6, 7)
All 500 replicates were merged together
(steps 8, 9)
Each SNP site was assigned a bin number according to the SNP's location
(step 10)
SD of all qualified minor allele frequency was obtained
(minor allele frequency was set as 'NA' if its non-missing allele number less than 2*N*0.5)
(step 11)
Standard deviation of 'SDs of each SNP site' by chromosomes
Counted the number of outlier SNP sites by bins
Box-plots of the SD in each chromosome
Dot plots of the SNP sites with higher SD
Pair-wise t-test analysis of mean SD
(step 12)
Different significant level was set for each of different sample sizes as above,
p= 0.05, 0.01, 0.005, 0.001, 0.0005 and 0.0001, respectively
Counted the number of significant replicates for each SNP site
Calculated the average number of significant false positives by bins
Set number of false positives was set as the outcome
Univariate and multivariate regression for finding predictors

SNP: single nucleotide polymorphism; SD: standard deviation
§ refers to each step of UNIX and R code in the appendix (steps 1, 2 are execute commands, not shown here)

**Figure 2. Operational flowchart of data analysis in UNIX and R**

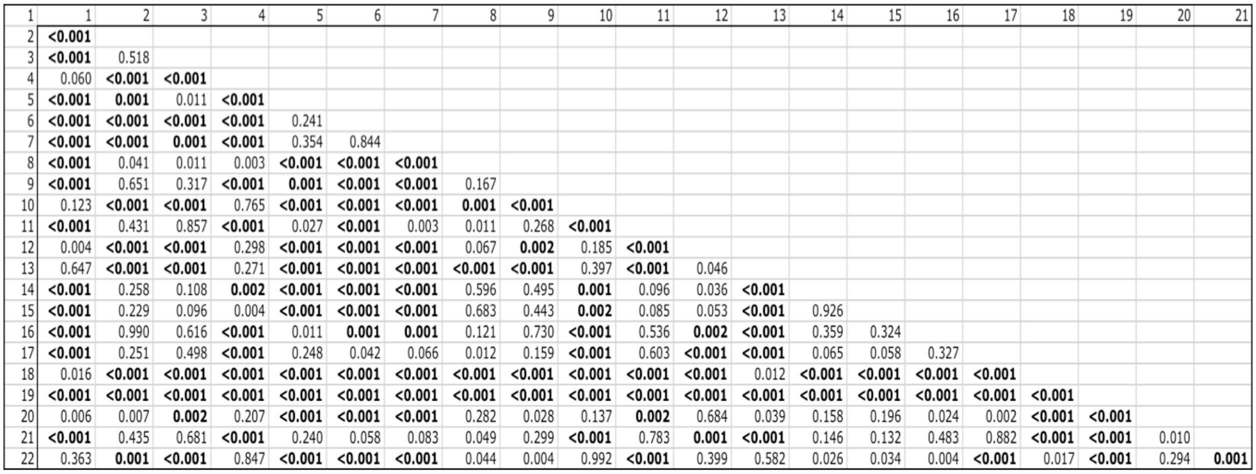
(a)



(b)

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.06314 | 0.06416 | 0.06424 | 0.06341 | 0.0646 | 0.06477 | 0.06474 | 0.06387 | 0.06409 | 0.06336 | 0.06427 |
| 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| 0.06357 | 0.06321 | 0.06396 | 0.06394 | 0.06415 | 0.06437 | 0.06269 | 0.06561 | 0.06366 | 0.06434 | 0.06336 |

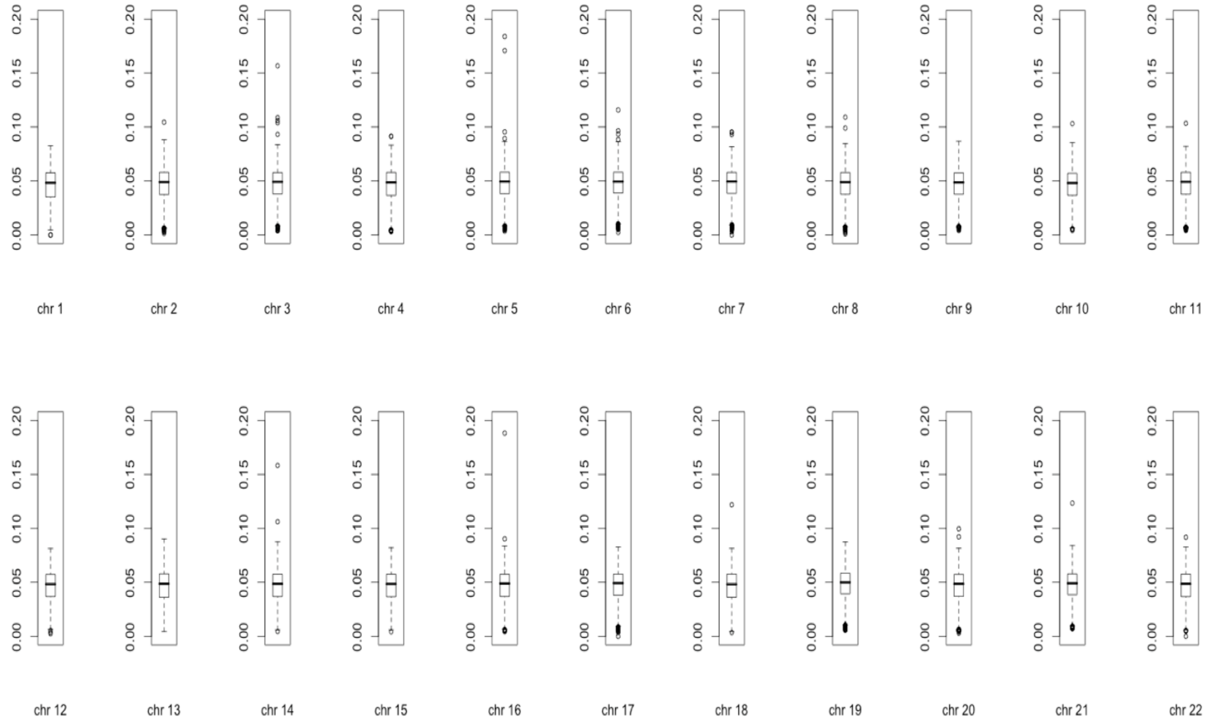
(c)



| 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | **<0.001** | | | | | | | | | | | | | | | | | | | | |
| 3 | **<0.001** | 0.518 | | | | | | | | | | | | | | | | | | | |
| 4 | 0.060 | **<0.001** | **<0.001** | | | | | | | | | | | | | | | | | | |
| 5 | **<0.001** | **0.001** | 0.011 | **<0.001** | | | | | | | | | | | | | | | | | |
| 6 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.241 | | | | | | | | | | | | | | | | |
| 7 | **<0.001** | **<0.001** | **0.001** | **<0.001** | 0.354 | 0.844 | | | | | | | | | | | | | | | |
| 8 | **<0.001** | 0.041 | 0.011 | 0.003 | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | | | |
| 9 | **<0.001** | 0.651 | 0.317 | **<0.001** | **0.001** | **<0.001** | **<0.001** | 0.167 | | | | | | | | | | | | | |
| 10 | 0.123 | **<0.001** | **<0.001** | 0.765 | **<0.001** | **<0.001** | **<0.001** | **0.001** | **<0.001** | | | | | | | | | | | | |
| 11 | **<0.001** | 0.431 | 0.857 | **<0.001** | 0.027 | **<0.001** | 0.003 | 0.011 | 0.268 | **<0.001** | | | | | | | | | | | |
| 12 | 0.004 | **<0.001** | **<0.001** | 0.298 | **<0.001** | **<0.001** | **<0.001** | 0.067 | **0.002** | 0.185 | **<0.001** | | | | | | | | | | |
| 13 | 0.647 | **<0.001** | **<0.001** | 0.271 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.397 | **<0.001** | 0.046 | | | | | | | | | |
| 14 | **<0.001** | 0.258 | 0.108 | **0.002** | **<0.001** | **<0.001** | **<0.001** | 0.596 | 0.495 | **0.001** | 0.096 | 0.036 | **<0.001** | | | | | | | | |
| 15 | **<0.001** | 0.229 | 0.096 | 0.004 | **<0.001** | **<0.001** | **<0.001** | 0.683 | 0.443 | **0.002** | 0.085 | 0.053 | **<0.001** | 0.926 | | | | | | | |
| 16 | **<0.001** | 0.990 | 0.616 | **<0.001** | 0.011 | **0.001** | **0.001** | 0.121 | 0.730 | **<0.001** | 0.536 | **0.002** | **<0.001** | 0.359 | 0.324 | | | | | | |
| 17 | **<0.001** | 0.251 | 0.498 | **<0.001** | 0.248 | 0.042 | 0.066 | 0.012 | 0.159 | **<0.001** | 0.603 | **<0.001** | **<0.001** | 0.065 | 0.058 | 0.327 | | | | | |
| 18 | 0.016 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.012 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | | |
| 19 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | |
| 20 | 0.006 | 0.007 | **0.002** | 0.207 | **<0.001** | **<0.001** | **<0.001** | 0.282 | 0.028 | 0.137 | **0.002** | 0.684 | 0.039 | 0.158 | 0.196 | 0.024 | 0.002 | **<0.001** | **<0.001** | | |
| 21 | **<0.001** | 0.435 | 0.681 | **<0.001** | 0.240 | 0.058 | 0.083 | 0.049 | 0.299 | **<0.001** | 0.783 | **0.001** | **<0.001** | 0.146 | 0.132 | 0.483 | 0.882 | **<0.001** | **<0.001** | 0.010 | |
| 22 | 0.363 | **0.001** | **<0.001** | 0.847 | **<0.001** | **<0.001** | **<0.001** | 0.044 | 0.004 | 0.992 | **<0.001** | 0.399 | 0.582 | 0.026 | 0.034 | 0.004 | **<0.001** | 0.017 | **<0.001** | 0.294 | **0.001** |

**Figure 3-1. N=25, replicates=500 (a) Standard deviation of minor allele frequencies for each chromosome; (b) Average standard deviation of minor allele frequency for each chromosome; (c) P values of pair-wise t-test comparing chromosome-specific means of standard deviation**

(a)



(b)

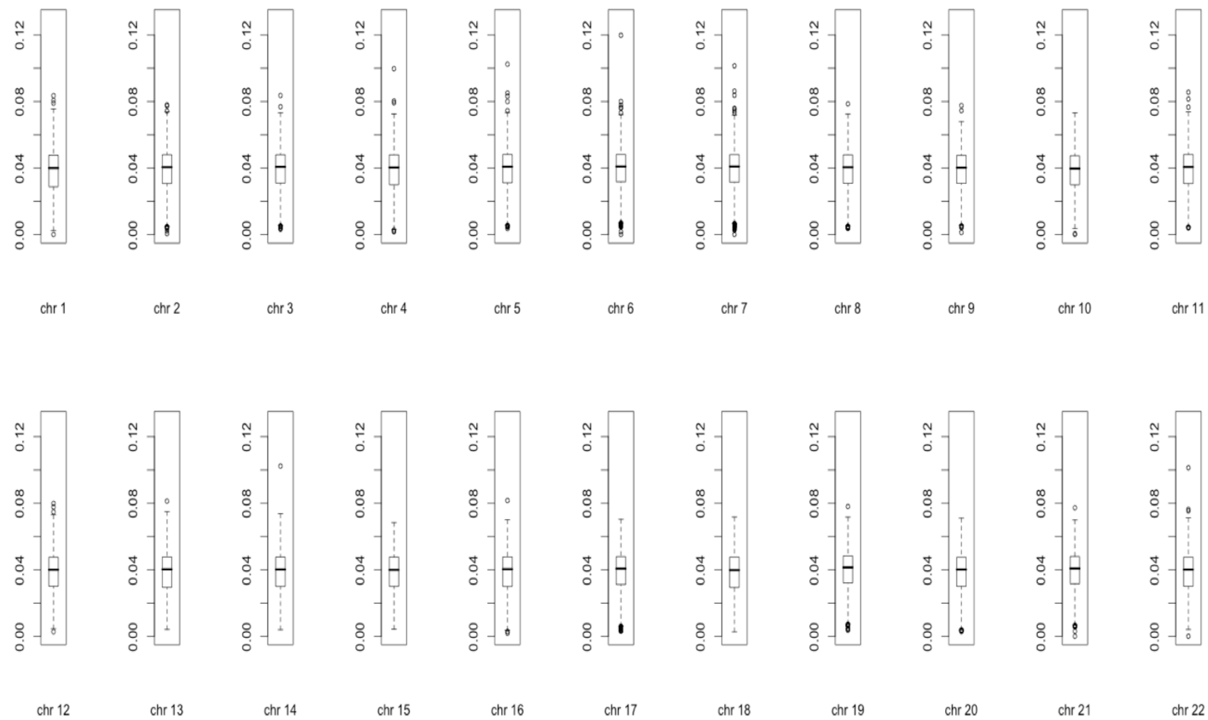
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.04562 | 0.04657 | 0.04674 | 0.04601 | 0.0471 | 0.04728 | 0.04706 | 0.04644 | 0.04657 | 0.04594 | 0.04668 |
| 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| 0.04615 | 0.04605 | 0.04624 | 0.0461 | 0.04626 | 0.04663 | 0.04554 | 0.0476 | 0.04604 | 0.04697 | 0.04613 |

(c)

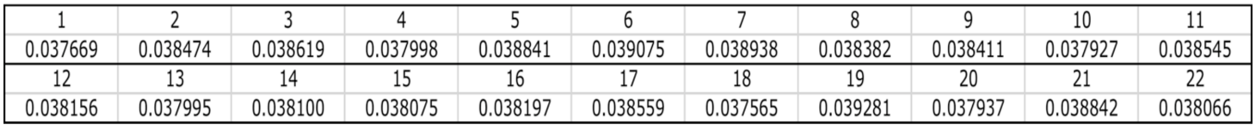
| 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | **<0.001** | | | | | | | | | | | | | | | | | | | | |
| 3 | **<0.001** | 0.120 | | | | | | | | | | | | | | | | | | | |
| 4 | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | | | | | | | |
| 5 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | | | | | | |
| 6 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.100 | | | | | | | | | | | | | | | | |
| 7 | **<0.001** | **<0.001** | 0.010 | **<0.001** | 0.730 | 0.050 | | | | | | | | | | | | | | | |
| 8 | **<0.001** | 0.220 | 0.010 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | | | |
| 9 | **<0.001** | 0.980 | 0.160 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.290 | | | | | | | | | | | | | |
| 10 | 0.010 | **<0.001** | **<0.001** | 0.540 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | |
| 11 | **<0.001** | 0.340 | 0.620 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.050 | 0.390 | **<0.001** | | | | | | | | | | | |
| 12 | **<0.001** | **<0.001** | **<0.001** | 0.280 | **<0.001** | **<0.001** | **<0.001** | 0.020 | **<0.001** | 0.090 | **<0.001** | | | | | | | | | | |
| 13 | **<0.001** | **<0.001** | **<0.001** | 0.810 | **<0.001** | **<0.001** | **<0.001** | 0.010 | **<0.001** | 0.430 | **<0.001** | 0.480 | | | | | | | | | |
| 14 | **<0.001** | 0.010 | **<0.001** | 0.120 | **<0.001** | **<0.001** | **<0.001** | 0.170 | 0.020 | 0.040 | **<0.001** | 0.530 | 0.230 | | | | | | | | |
| 15 | **<0.001** | **<0.001** | **<0.001** | 0.570 | **<0.001** | **<0.001** | **<0.001** | 0.020 | **<0.001** | 0.280 | **<0.001** | 0.740 | 0.760 | 0.390 | | | | | | | |
| 16 | **<0.001** | 0.020 | **<0.001** | 0.100 | **<0.001** | **<0.001** | **<0.001** | 0.210 | 0.030 | 0.030 | **<0.001** | 0.460 | 0.190 | 0.930 | 0.340 | | | | | | |
| 17 | **<0.001** | 0.710 | 0.480 | **<0.001** | **<0.001** | **<0.001** | 0.010 | 0.220 | 0.710 | 0.000 | 0.750 | **<0.001** | **<0.001** | 0.030 | **<0.001** | 0.030 | | | | | |
| 18 | 0.560 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.010 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | | |
| 19 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.010 | 0.080 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | |
| 20 | 0.010 | **<0.001** | **<0.001** | 0.890 | **<0.001** | **<0.001** | **<0.001** | 0.010 | **<0.001** | 0.530 | **<0.001** | 0.480 | 0.950 | 0.240 | 0.730 | 0.210 | **<0.001** | 0.010 | **<0.001** | | |
| 21 | **<0.001** | 0.020 | 0.210 | **<0.001** | 0.460 | 0.070 | 0.610 | **<0.001** | 0.030 | **<0.001** | 0.120 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.110 | **<0.001** | 0.010 | **<0.001** | |
| 22 | 0.010 | 0.020 | **<0.001** | 0.550 | **<0.001** | **<0.001** | **<0.001** | 0.130 | 0.030 | 0.330 | 0.010 | 0.940 | 0.690 | 0.610 | 0.870 | 0.570 | 0.020 | 0.010 | **<0.001** | 0.660 | **<0.001** |

**Figure 3-2. N=50, replicates=500 (a) Standard deviation of minor allele frequencies for each chromosome; (b) Average standard deviation of minor allele frequency for each chromosome; (c) P values of pair-wise t-test comparing chromosome-specific means of standard deviation**

(a)

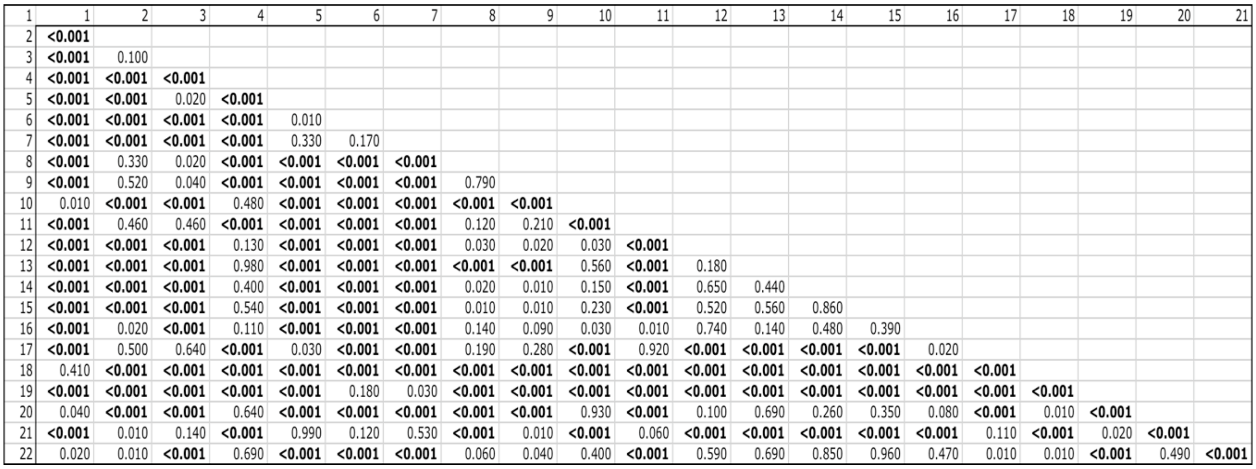


(b)



| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.037669 | 0.038474 | 0.038619 | 0.037998 | 0.038841 | 0.039075 | 0.038938 | 0.038382 | 0.038411 | 0.037927 | 0.038545 |
| 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| 0.038156 | 0.037995 | 0.038100 | 0.038075 | 0.038197 | 0.038559 | 0.037565 | 0.039281 | 0.037937 | 0.038842 | 0.038066 |

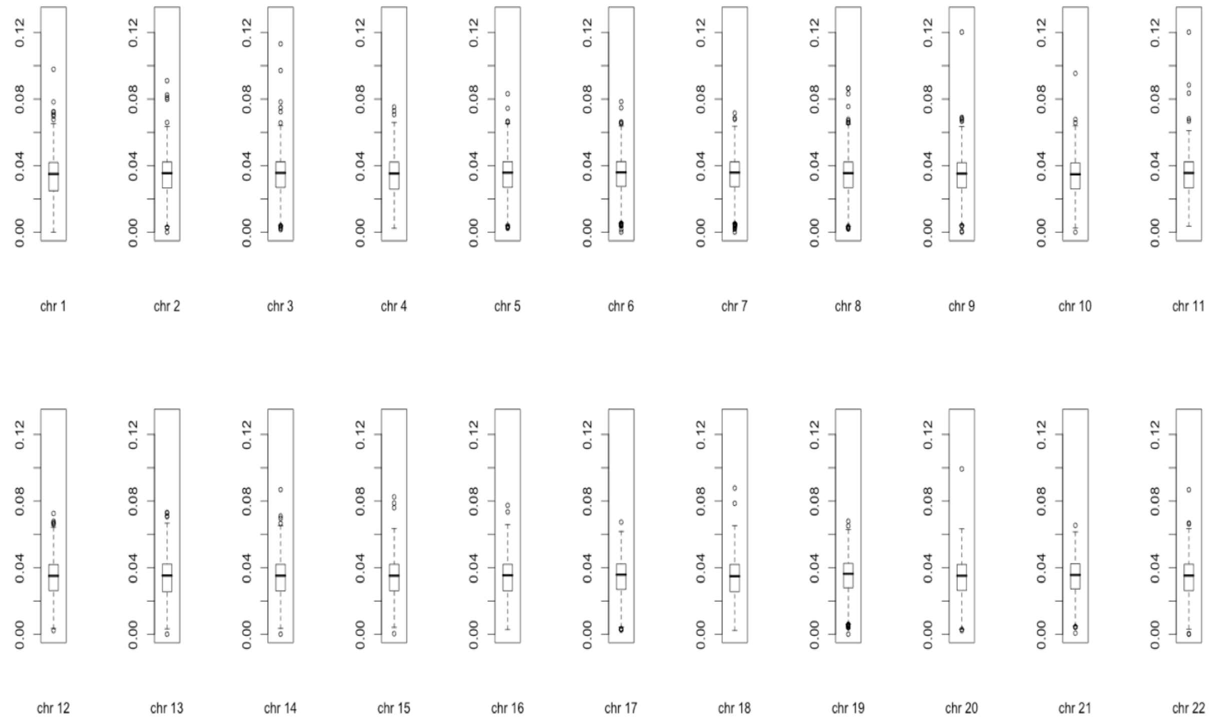
(c)



| 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | **<0.001** | | | | | | | | | | | | | | | | | | | | |
| 3 | **<0.001** | 0.100 | | | | | | | | | | | | | | | | | | | |
| 4 | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | | | | | | | |
| 5 | **<0.001** | **<0.001** | 0.020 | **<0.001** | | | | | | | | | | | | | | | | | |
| 6 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.010 | | | | | | | | | | | | | | | | |
| 7 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.330 | 0.170 | | | | | | | | | | | | | | | |
| 8 | **<0.001** | 0.330 | 0.020 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | | | |
| 9 | **<0.001** | 0.520 | 0.040 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.790 | | | | | | | | | | | | | |
| 10 | 0.010 | **<0.001** | **<0.001** | 0.480 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | |
| 11 | **<0.001** | 0.460 | 0.460 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.120 | 0.210 | **<0.001** | | | | | | | | | | | |
| 12 | **<0.001** | **<0.001** | **<0.001** | 0.130 | **<0.001** | **<0.001** | **<0.001** | 0.030 | 0.020 | 0.030 | **<0.001** | | | | | | | | | | |
| 13 | **<0.001** | **<0.001** | **<0.001** | 0.980 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.560 | **<0.001** | 0.180 | | | | | | | | | |
| 14 | **<0.001** | **<0.001** | **<0.001** | 0.400 | **<0.001** | **<0.001** | **<0.001** | 0.020 | 0.010 | 0.150 | **<0.001** | 0.650 | 0.440 | | | | | | | | |
| 15 | **<0.001** | **<0.001** | **<0.001** | 0.540 | **<0.001** | **<0.001** | **<0.001** | 0.010 | 0.010 | 0.230 | **<0.001** | 0.520 | 0.560 | 0.860 | | | | | | | |
| 16 | **<0.001** | 0.020 | **<0.001** | 0.110 | **<0.001** | **<0.001** | **<0.001** | 0.140 | 0.090 | 0.030 | 0.010 | 0.740 | 0.140 | 0.480 | 0.390 | | | | | | |
| 17 | **<0.001** | 0.500 | 0.640 | **<0.001** | 0.030 | **<0.001** | **<0.001** | 0.190 | 0.280 | **<0.001** | 0.920 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.020 | | | | | |
| 18 | 0.410 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | | |
| 19 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.180 | 0.030 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | |
| 20 | 0.040 | **<0.001** | **<0.001** | 0.640 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.930 | **<0.001** | 0.100 | 0.690 | 0.260 | 0.350 | 0.080 | **<0.001** | 0.010 | **<0.001** | | |
| 21 | **<0.001** | 0.010 | 0.140 | **<0.001** | 0.990 | 0.120 | 0.530 | **<0.001** | 0.010 | **<0.001** | 0.060 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.110 | **<0.001** | 0.020 | **<0.001** | |
| 22 | 0.020 | 0.010 | **<0.001** | 0.690 | **<0.001** | **<0.001** | **<0.001** | 0.060 | 0.040 | 0.400 | **<0.001** | 0.590 | 0.690 | 0.850 | 0.960 | 0.470 | 0.010 | 0.010 | **<0.001** | 0.490 | **<0.001** |

**Figure 3-3. N=75, replicates=500 (a) Standard deviation of minor allele frequencies for each chromosome; (b) Average standard deviation of minor allele frequency for each chromosome; (c) P values of pair-wise t-test comparing chromosome-specific means of standard deviation**

(a)



(b)

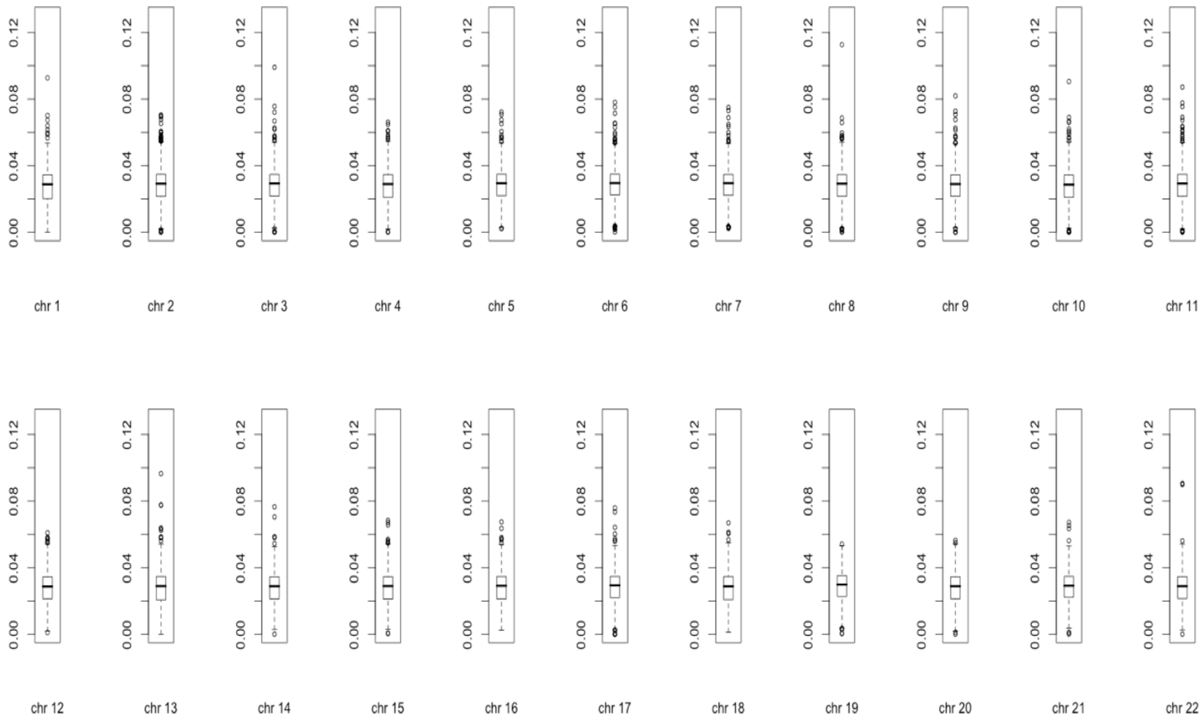
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.032918 | 0.033600 | 0.033820 | 0.033226 | 0.033989 | 0.034197 | 0.034045 | 0.033612 | 0.033599 | 0.033179 | 0.033674 |
| 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| 0.033360 | 0.033229 | 0.033371 | 0.033322 | 0.033426 | 0.033760 | 0.032898 | 0.034392 | 0.033299 | 0.033980 | 0.033345 |

(c)

| 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | **<0.001** | | | | | | | | | | | | | | | | | | | | |
| 3 | **<0.001** | 0.010 | | | | | | | | | | | | | | | | | | | |
| 4 | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | | | | | | | |
| 5 | **<0.001** | **<0.001** | 0.040 | **<0.001** | | | | | | | | | | | | | | | | | |
| 6 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.010 | | | | | | | | | | | | | | | | |
| 7 | **<0.001** | **<0.001** | 0.010 | **<0.001** | 0.530 | 0.090 | | | | | | | | | | | | | | | |
| 8 | **<0.001** | 0.900 | 0.020 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | | | |
| 9 | **<0.001** | 0.980 | 0.020 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.890 | | | | | | | | | | | | | |
| 10 | **<0.001** | **<0.001** | **<0.001** | 0.600 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | |
| 11 | **<0.001** | 0.390 | 0.100 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.510 | 0.440 | **<0.001** | | | | | | | | | | | |
| 12 | **<0.001** | 0.010 | **<0.001** | 0.150 | **<0.001** | **<0.001** | **<0.001** | 0.010 | 0.010 | 0.050 | **<0.001** | | | | | | | | | | |
| 13 | **<0.001** | **<0.001** | **<0.001** | 0.980 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.630 | **<0.001** | 0.220 | | | | | | | | | |
| 14 | **<0.001** | 0.020 | **<0.001** | 0.180 | **<0.001** | **<0.001** | **<0.001** | 0.030 | 0.040 | 0.080 | 0.010 | 0.920 | 0.240 | | | | | | | | |
| 15 | **<0.001** | 0.010 | **<0.001** | 0.390 | **<0.001** | **<0.001** | **<0.001** | 0.010 | 0.010 | 0.200 | **<0.001** | 0.730 | 0.450 | 0.690 | | | | | | | |
| 16 | **<0.001** | 0.090 | **<0.001** | 0.070 | **<0.001** | **<0.001** | **<0.001** | 0.090 | 0.120 | 0.020 | 0.030 | 0.560 | 0.110 | 0.660 | 0.410 | | | | | | |
| 17 | **<0.001** | 0.160 | 0.610 | **<0.001** | 0.050 | **<0.001** | 0.020 | 0.220 | 0.190 | **<0.001** | 0.480 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.010 | | | | | |
| 18 | 0.850 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.010 | **<0.001** | **<0.001** | 0.010 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | | |
| 19 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.160 | 0.010 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | |
| 20 | **<0.001** | 0.010 | **<0.001** | 0.530 | **<0.001** | **<0.001** | **<0.001** | 0.010 | 0.010 | 0.300 | **<0.001** | 0.610 | 0.580 | 0.580 | 0.870 | 0.340 | **<0.001** | **<0.001** | **<0.001** | | |
| 21 | **<0.001** | **<0.001** | 0.240 | **<0.001** | 0.950 | 0.110 | 0.640 | 0.010 | 0.010 | **<0.001** | 0.030 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.160 | **<0.001** | 0.020 | **<0.001** | |
| 22 | **<0.001** | 0.080 | **<0.001** | 0.430 | **<0.001** | **<0.001** | **<0.001** | 0.080 | 0.090 | 0.270 | 0.030 | 0.920 | 0.470 | 0.870 | 0.890 | 0.620 | 0.010 | 0.010 | **<0.001** | 0.790 | **<0.001** |

**Figure 3-4. N=100, replicates=500 (a) Standard deviation of minor allele frequencies for each chromosome; (b) Average standard deviation of minor allele frequency for each chromosome; (c) P values of pair-wise t-test comparing chromosome-specific means of standard deviation**

(a)



(b)

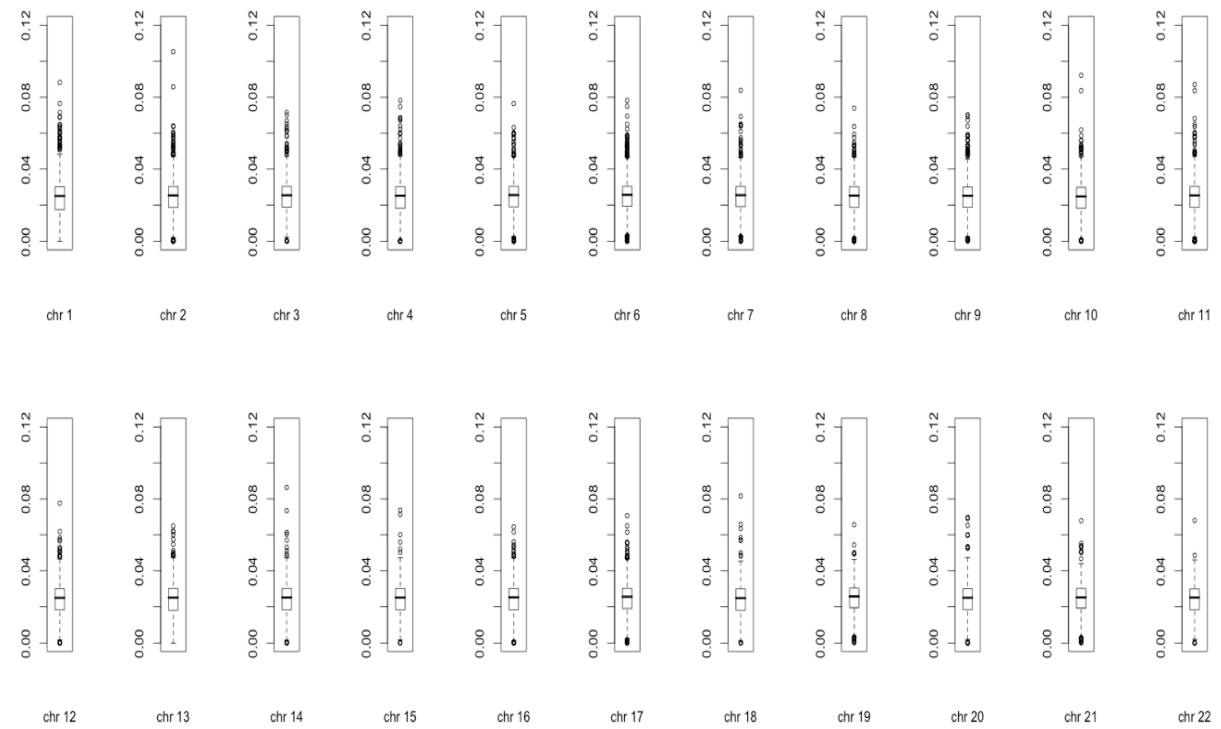
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.026963 | 0.027605 | 0.027703 | 0.027206 | 0.027871 | 0.028032 | 0.027913 | 0.027531 | 0.027555 | 0.027232 | 0.027607 |
| 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| 0.027310 | 0.027236 | 0.027323 | 0.027298 | 0.027407 | 0.027692 | 0.027023 | 0.028175 | 0.027258 | 0.027796 | 0.027366 |

(c)

| 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | **<0.001** | | | | | | | | | | | | | | | | | | | | |
| 3 | **<0.001** | 0.150 | | | | | | | | | | | | | | | | | | | |
| 4 | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | | | | | | | |
| 5 | **<0.001** | **<0.001** | 0.020 | **<0.001** | | | | | | | | | | | | | | | | | |
| 6 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.020 | | | | | | | | | | | | | | | | |
| 7 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.580 | 0.110 | | | | | | | | | | | | | | | |
| 8 | **<0.001** | 0.300 | 0.020 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | | | |
| 9 | **<0.001** | 0.500 | 0.050 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.770 | | | | | | | | | | | | | |
| 10 | **<0.001** | **<0.001** | **<0.001** | 0.730 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | |
| 11 | **<0.001** | 0.980 | 0.200 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.340 | 0.530 | **<0.001** | | | | | | | | | | | |
| 12 | **<0.001** | **<0.001** | **<0.001** | 0.180 | **<0.001** | **<0.001** | **<0.001** | 0.010 | **<0.001** | 0.320 | **<0.001** | | | | | | | | | | |
| 13 | **<0.001** | **<0.001** | **<0.001** | 0.730 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.970 | **<0.001** | 0.410 | | | | | | | | | |
| 14 | **<0.001** | **<0.001** | **<0.001** | 0.200 | **<0.001** | **<0.001** | **<0.001** | 0.020 | 0.010 | 0.320 | **<0.001** | 0.890 | 0.390 | | | | | | | | |
| 15 | **<0.001** | **<0.001** | **<0.001** | 0.330 | **<0.001** | **<0.001** | **<0.001** | 0.010 | 0.010 | 0.480 | **<0.001** | 0.900 | 0.550 | 0.810 | | | | | | | |
| 16 | **<0.001** | 0.020 | **<0.001** | 0.030 | **<0.001** | **<0.001** | **<0.001** | 0.180 | 0.120 | 0.060 | 0.030 | 0.310 | 0.100 | 0.430 | 0.310 | | | | | | |
| 17 | **<0.001** | 0.370 | 0.910 | **<0.001** | 0.070 | **<0.001** | 0.030 | 0.120 | 0.180 | **<0.001** | 0.410 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.010 | | | | | |
| 18 | 0.520 | **<0.001** | **<0.001** | 0.060 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.030 | **<0.001** | **<0.001** | 0.040 | 0.010 | 0.010 | **<0.001** | **<0.001** | | | | |
| 19 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.010 | 0.220 | 0.030 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | |
| 20 | **<0.001** | **<0.001** | **<0.001** | 0.600 | **<0.001** | **<0.001** | **<0.001** | 0.010 | **<0.001** | 0.790 | **<0.001** | 0.600 | 0.840 | 0.550 | 0.720 | 0.180 | **<0.001** | 0.040 | **<0.001** | | |
| 21 | **<0.001** | 0.090 | 0.410 | **<0.001** | 0.510 | 0.040 | 0.310 | 0.030 | 0.040 | **<0.001** | 0.110 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.440 | **<0.001** | 0.010 | **<0.001** | |
| 22 | **<0.001** | 0.050 | 0.010 | 0.210 | **<0.001** | **<0.001** | **<0.001** | 0.200 | 0.130 | 0.290 | 0.060 | 0.660 | 0.330 | 0.750 | 0.620 | 0.770 | 0.020 | 0.010 | **<0.001** | 0.440 | 0.010 |

**Figure 3-5. N=150, replicates=500 (a) Standard deviation of minor allele frequencies for each chromosome; (b) Average standard deviation of minor allele frequency for each chromosome; (c) P values of pair-wise t-test comparing chromosome-specific means of standard deviation**

(a)



(b)

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.023642 | 0.024153 | 0.024279 | 0.023855 | 0.024379 | 0.024541 | 0.024403 | 0.024041 | 0.024128 | 0.023787 | 0.024168 |
| 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| 0.023912 | 0.023839 | 0.023972 | 0.023918 | 0.023957 | 0.024233 | 0.023565 | 0.024576 | 0.023864 | 0.024283 | 0.023813 |

(c)

| 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | **<0.001** | | | | | | | | | | | | | | | | | | | | |
| 3 | **<0.001** | 0.034 | | | | | | | | | | | | | | | | | | | |
| 4 | **0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | | | | | | | |
| 5 | **<0.001** | **<0.001** | 0.108 | **<0.001** | | | | | | | | | | | | | | | | | |
| 6 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.011 | | | | | | | | | | | | | | | | |
| 7 | **<0.001** | **<0.001** | 0.061 | **<0.001** | 0.720 | 0.039 | | | | | | | | | | | | | | | |
| 8 | **<0.001** | 0.073 | **<0.001** | 0.006 | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | | | |
| 9 | **<0.001** | 0.701 | 0.026 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.220 | | | | | | | | | | | | | |
| 10 | 0.029 | **<0.001** | **<0.001** | 0.313 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | | | | | | | | | | |
| 11 | **<0.001** | 0.812 | 0.098 | **<0.001** | 0.002 | **<0.001** | **0.001** | 0.069 | 0.577 | **<0.001** | | | | | | | | | | | |
| 12 | **<0.001** | **<0.001** | **<0.001** | 0.412 | **<0.001** | **<0.001** | **<0.001** | 0.068 | 0.003 | 0.074 | **<0.001** | | | | | | | | | | |
| 13 | 0.010 | **<0.001** | **<0.001** | 0.839 | **<0.001** | **<0.001** | **<0.001** | 0.010 | **<0.001** | 0.502 | **<0.001** | 0.358 | | | | | | | | | |
| 14 | **<0.001** | 0.016 | **<0.001** | 0.146 | **<0.001** | **<0.001** | **<0.001** | 0.397 | 0.057 | 0.021 | 0.017 | 0.463 | 0.135 | | | | | | | | |
| 15 | **0.001** | 0.002 | **<0.001** | 0.450 | **<0.001** | **<0.001** | **<0.001** | 0.140 | 0.012 | 0.113 | 0.003 | 0.946 | 0.389 | 0.559 | | | | | | | |
| 16 | **<0.001** | 0.011 | **<0.001** | 0.207 | **<0.001** | **<0.001** | **<0.001** | 0.309 | 0.040 | 0.037 | 0.011 | 0.588 | 0.192 | 0.871 | 0.676 | | | | | | |
| 17 | **<0.001** | 0.345 | 0.596 | **<0.001** | 0.089 | **<0.001** | 0.057 | 0.032 | 0.247 | **<0.001** | 0.474 | **<0.001** | **<0.001** | 0.008 | 0.002 | 0.006 | | | | | |
| 18 | 0.347 | **<0.001** | **<0.001** | **0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.008 | **<0.001** | **<0.001** | 0.003 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | | | | |
| 19 | **<0.001** | **<0.001** | 0.004 | **<0.001** | 0.056 | 0.729 | 0.099 | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | **<0.001** | 0.004 | **<0.001** | | | |
| 20 | 0.010 | **<0.001** | **<0.001** | 0.914 | **<0.001** | **<0.001** | **<0.001** | 0.043 | 0.003 | 0.373 | **0.001** | 0.589 | 0.792 | 0.265 | 0.589 | 0.343 | **<0.001** | 0.003 | **<0.001** | | |
| 21 | **<0.001** | 0.190 | 0.969 | **<0.001** | 0.339 | 0.011 | 0.243 | 0.019 | 0.137 | **<0.001** | 0.269 | **<0.001** | **<0.001** | 0.005 | **0.001** | 0.004 | 0.672 | **<0.001** | 0.023 | **<0.001** | |
| 22 | 0.131 | **0.001** | **<0.001** | 0.705 | **<0.001** | **<0.001** | **<0.001** | 0.041 | 0.005 | 0.813 | **0.002** | 0.373 | 0.826 | 0.181 | 0.386 | 0.231 | **0.001** | 0.042 | **<0.001** | 0.679 | **0.001** |

**Figure 3-6. N=200, replicates=500 (a) Standard deviation of minor allele frequencies for each chromosome; (b) Average standard deviation of minor allele frequency for each chromosome; (c) P values of pair-wise t-test comparing chromosome-specific means of standard deviation**

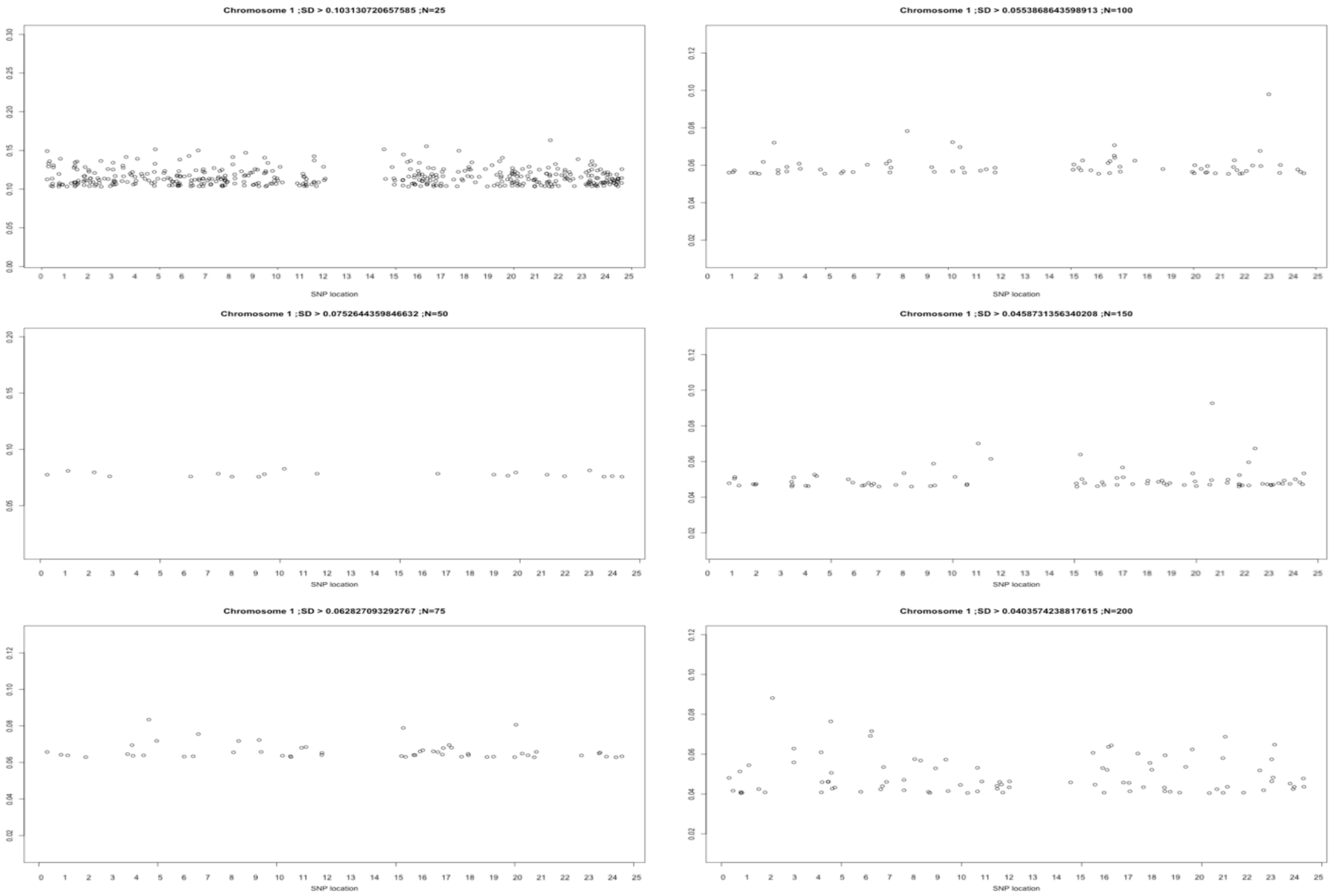


**Figure 4-1. Dot plots of outliers and their locations on chromosome 1 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

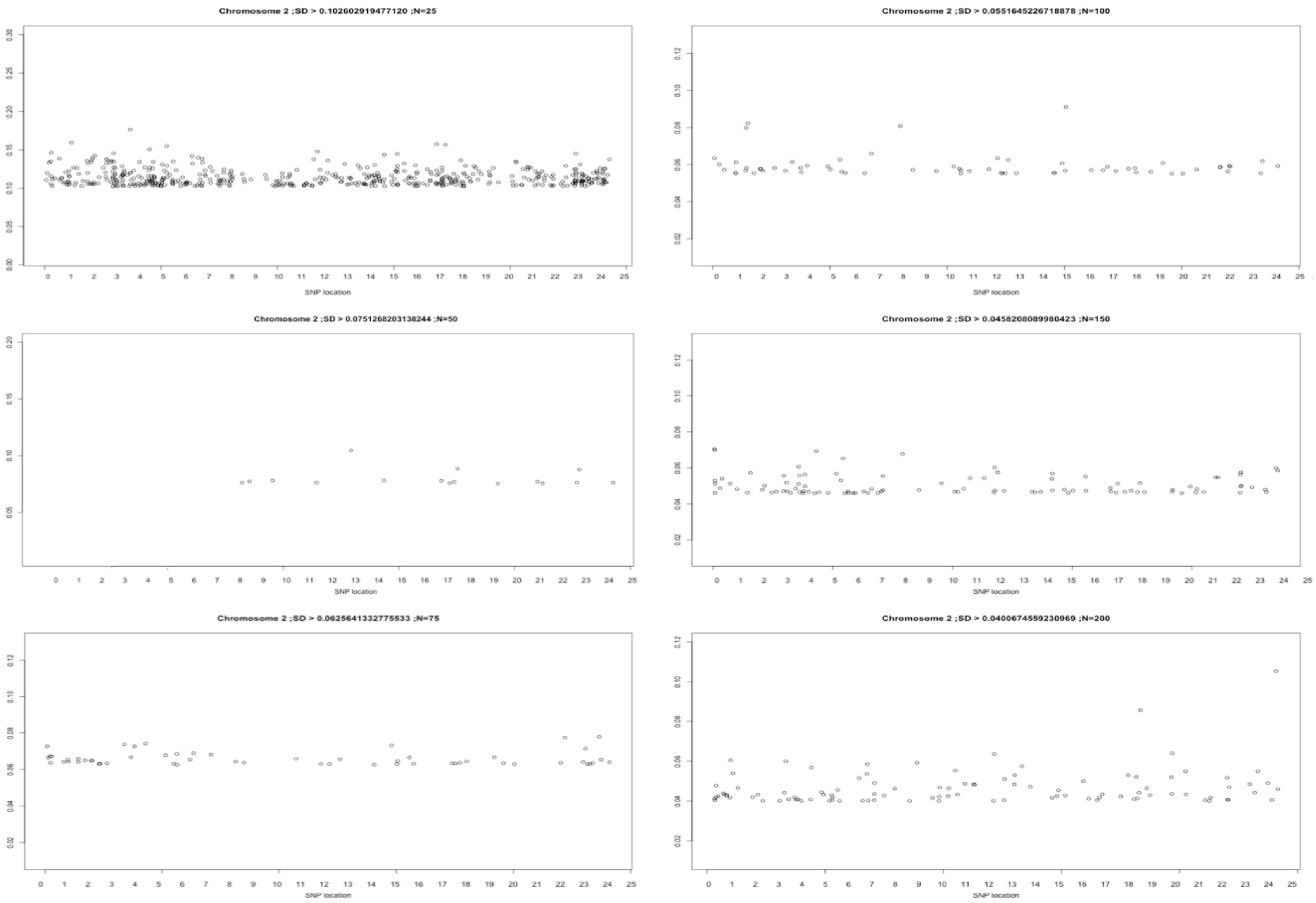


**Figure 4-2. Dot plots of outliers and their locations on chromosome 2 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

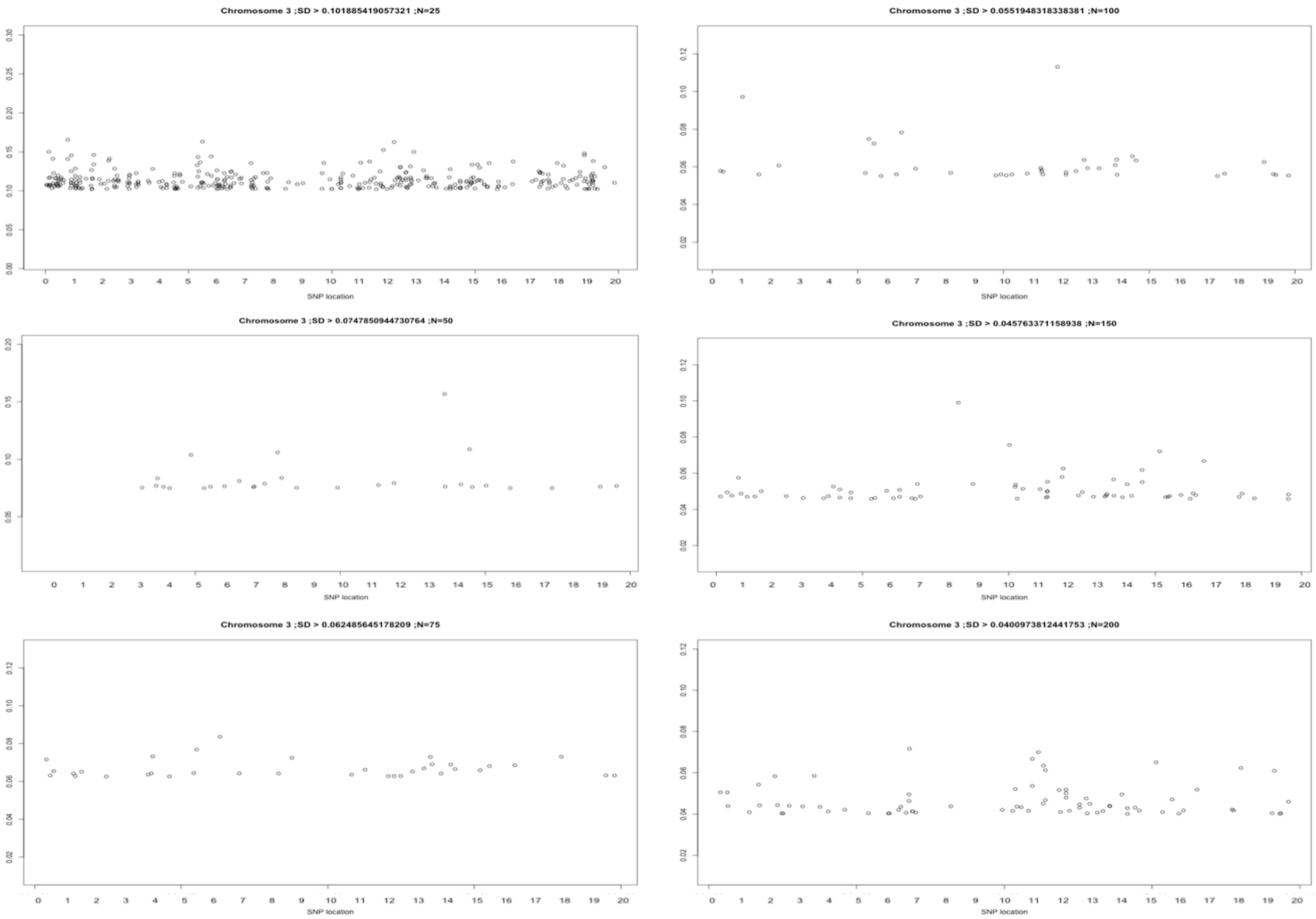


**Figure 4-3. Dot plots of outliers and their locations on chromosome 3 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

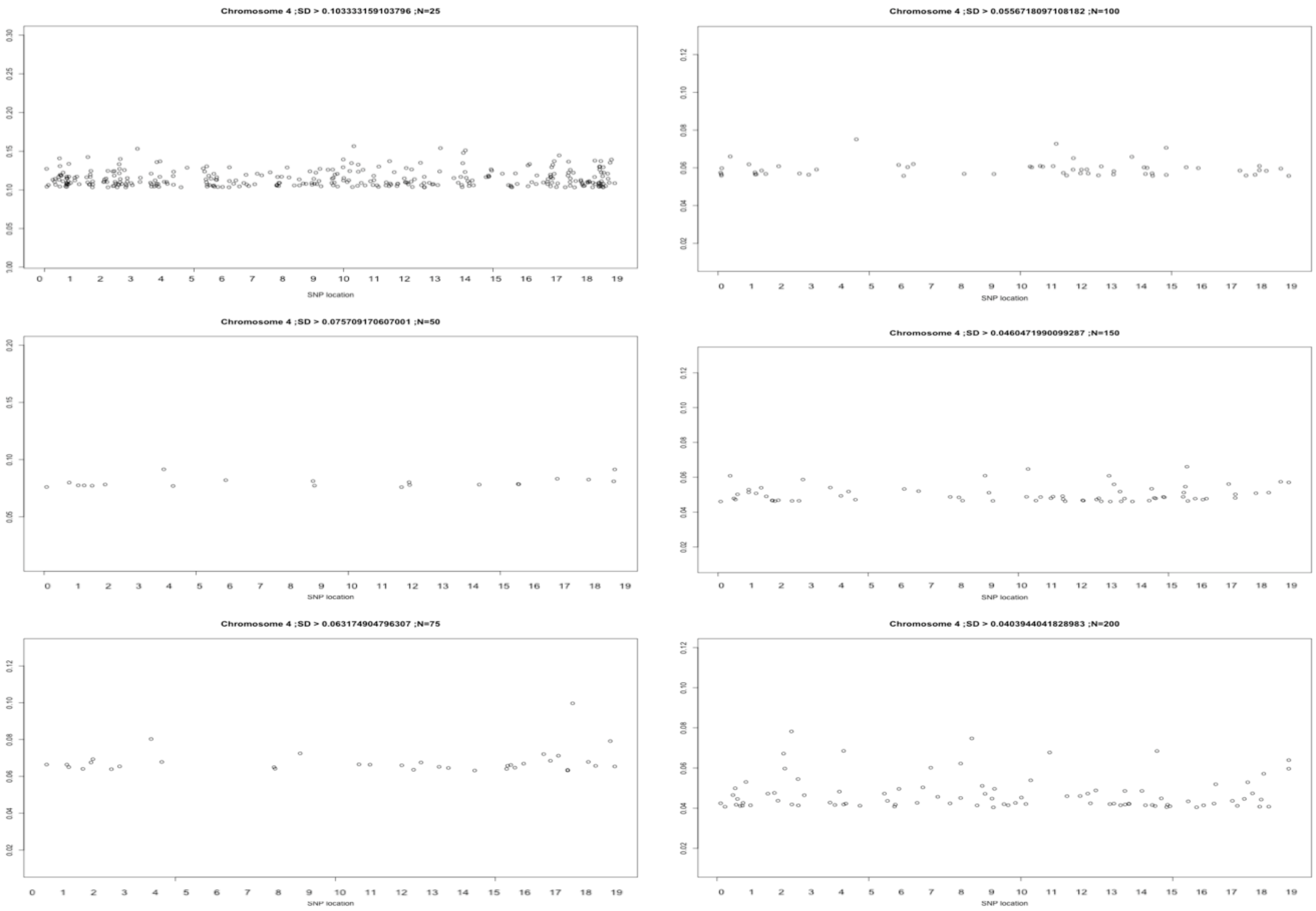


**Figure 4-4. Dot plots of outliers and their locations on chromosome 4 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

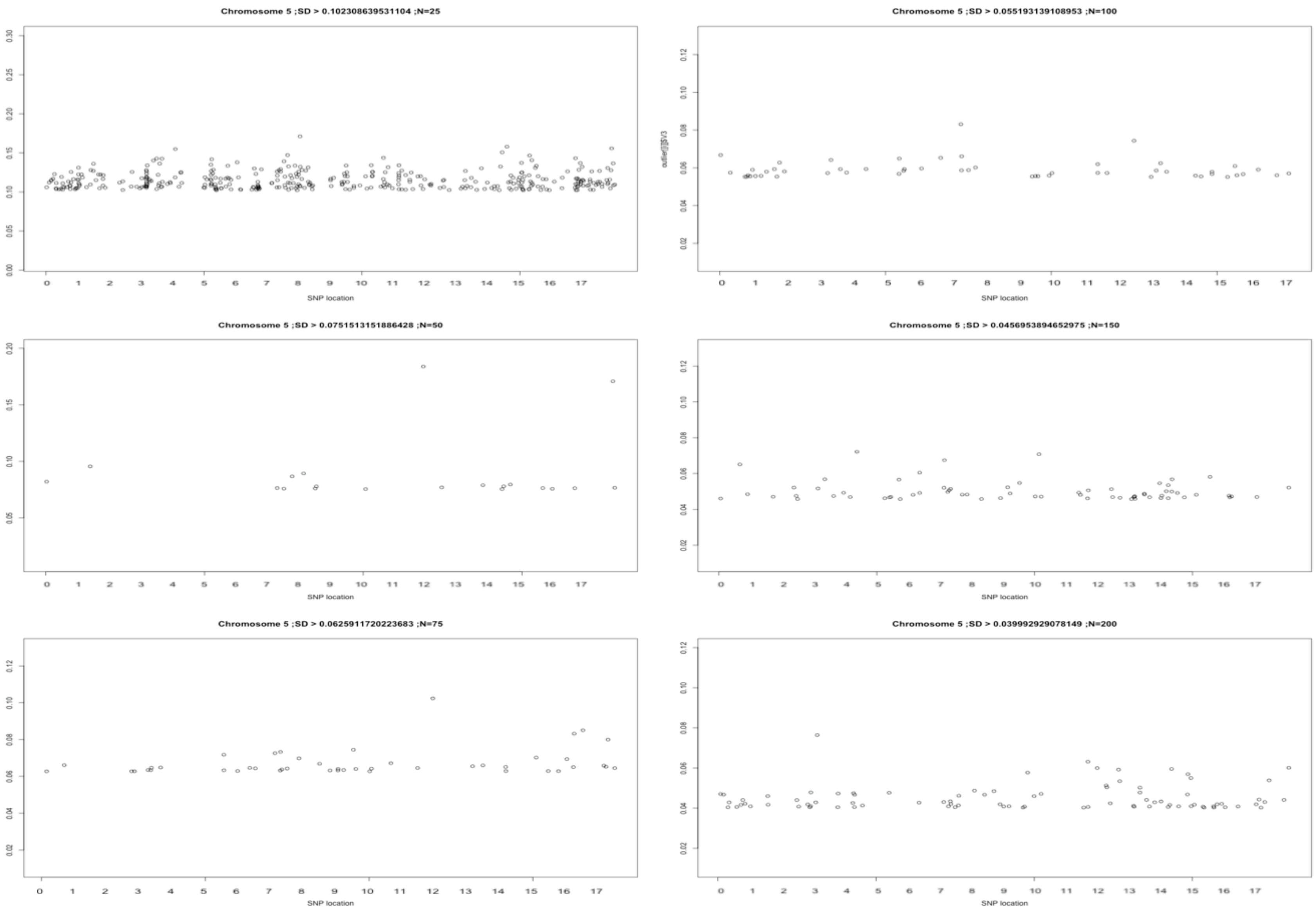


**Figure 4-5. Dot plots of outliers and their locations on chromosome 5 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

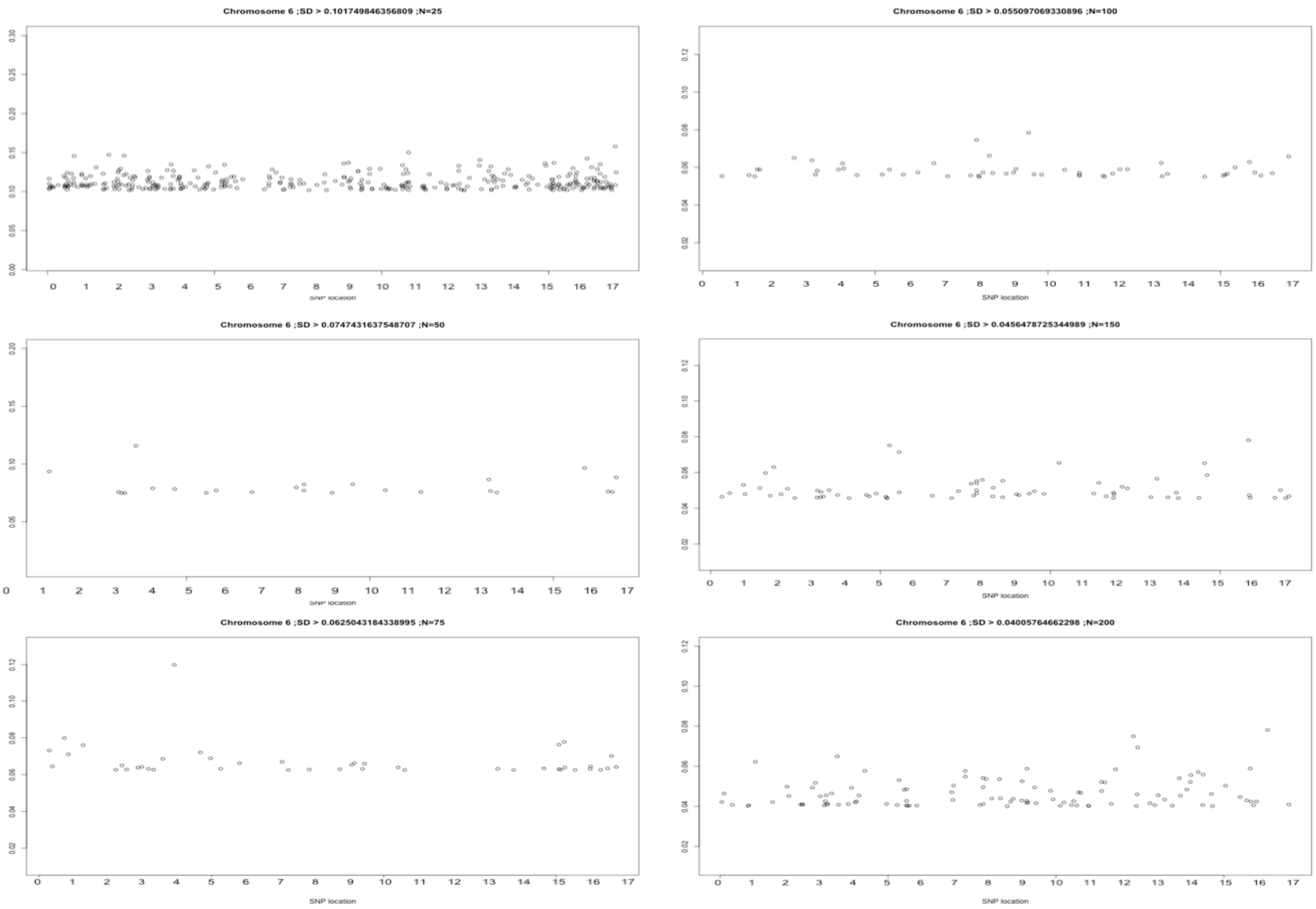


**Figure 4-6. Dot plots of outliers and their locations on chromosome 6 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

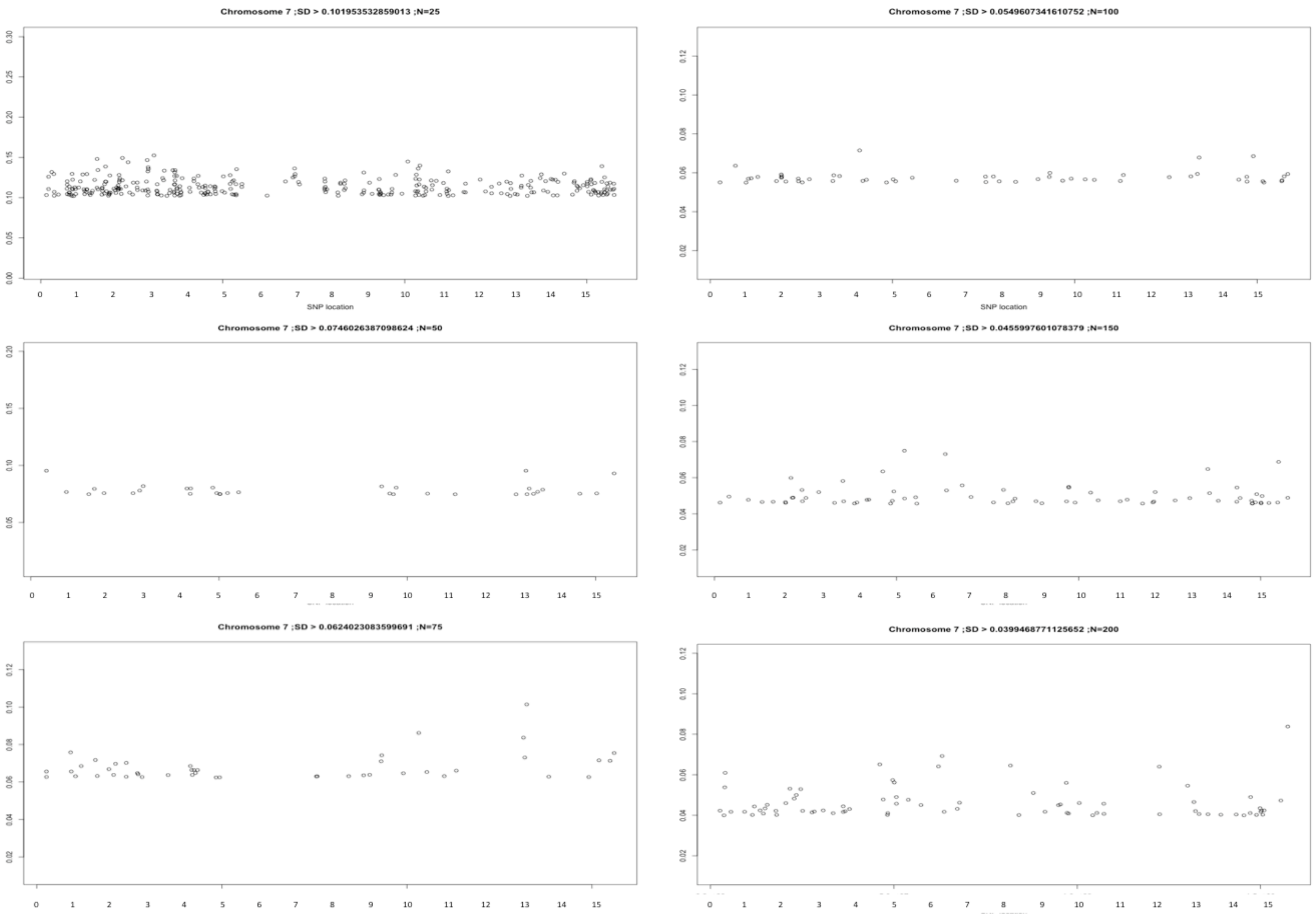


**Figure 4-7. Dot plots of outliers and their locations on chromosome 7 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

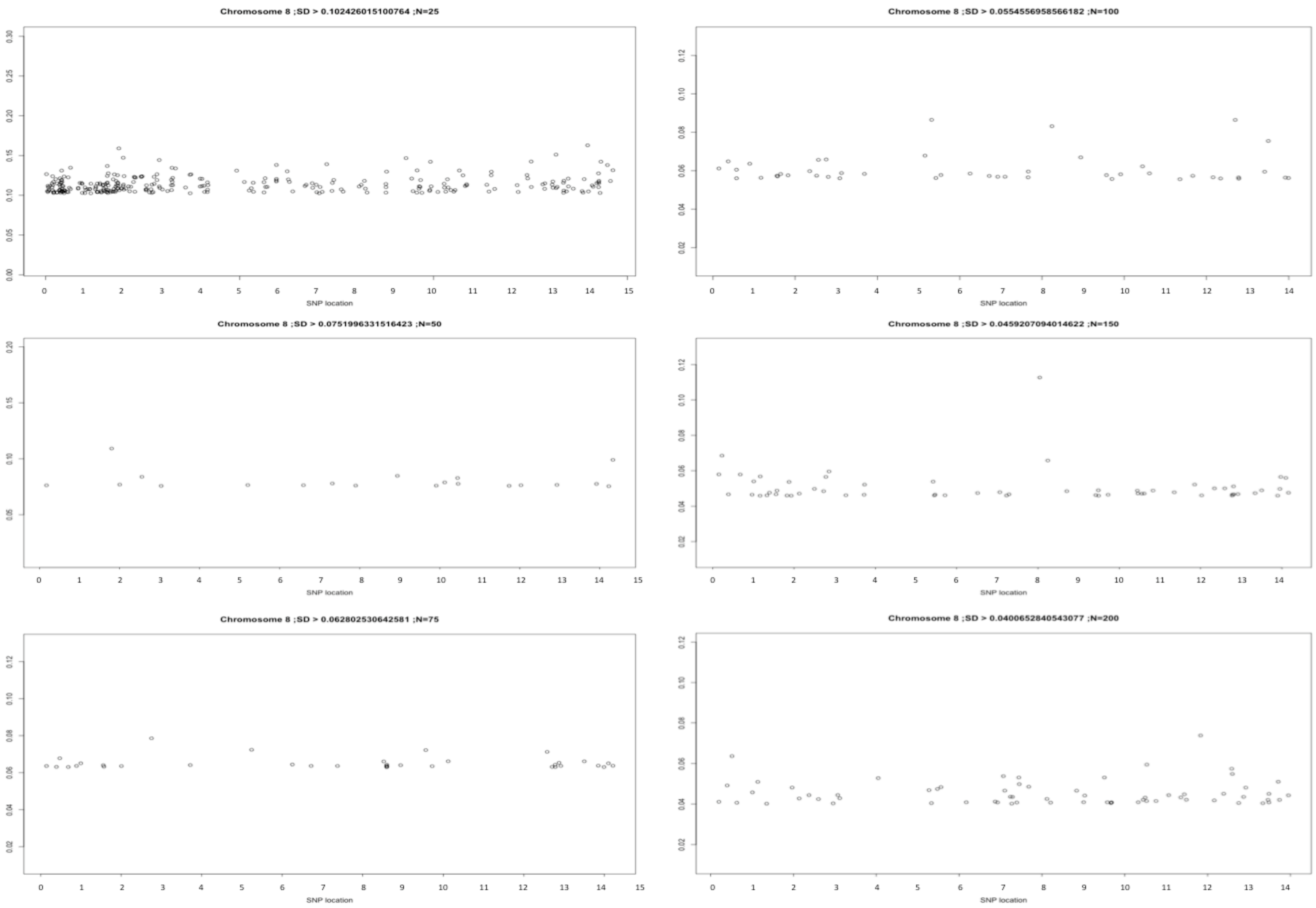


**Figure 4-8. Dot plots of outliers and their locations on chromosome 8 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

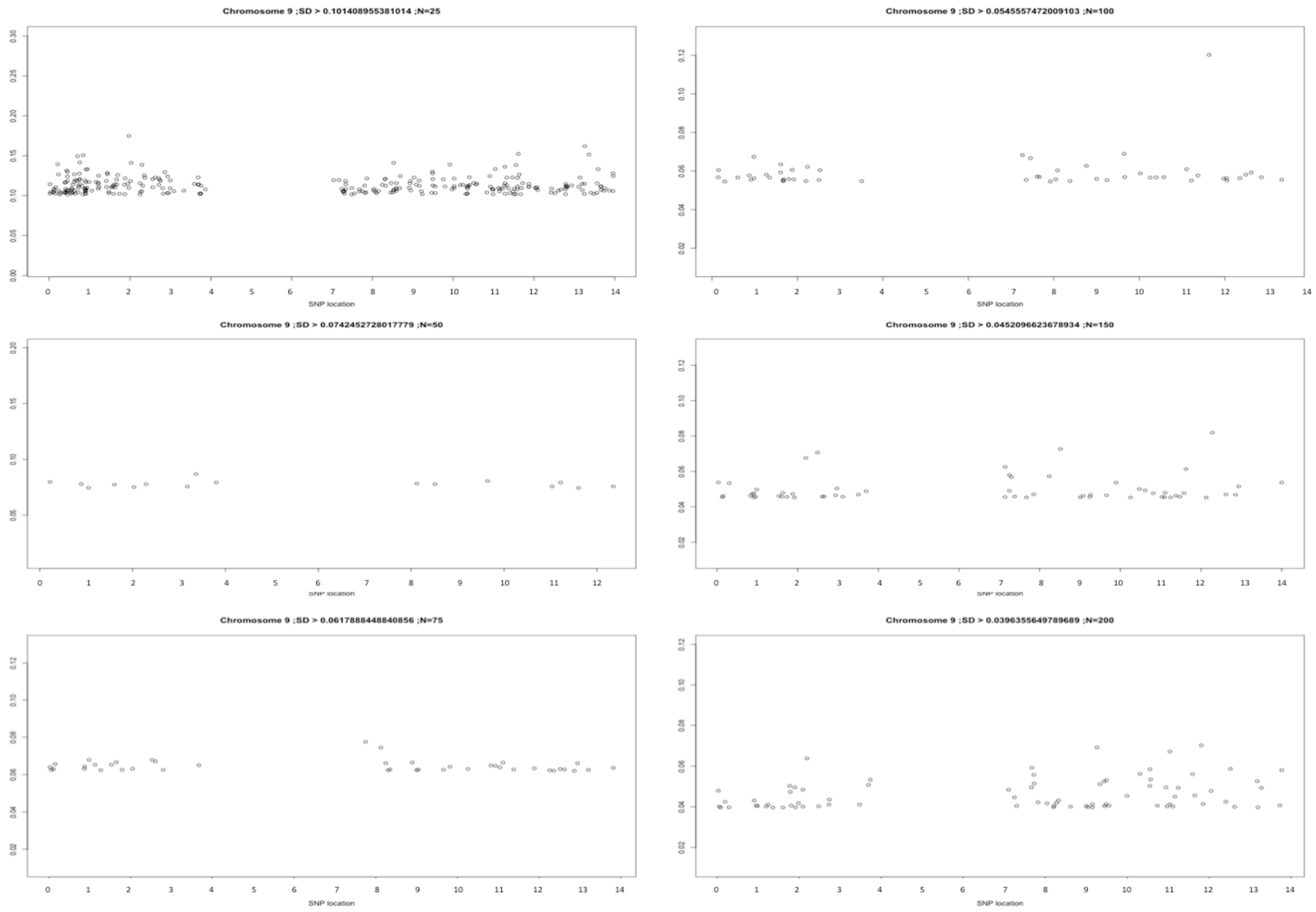


**Figure 4-9. Dot plots of outliers and their locations on chromosome 9 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

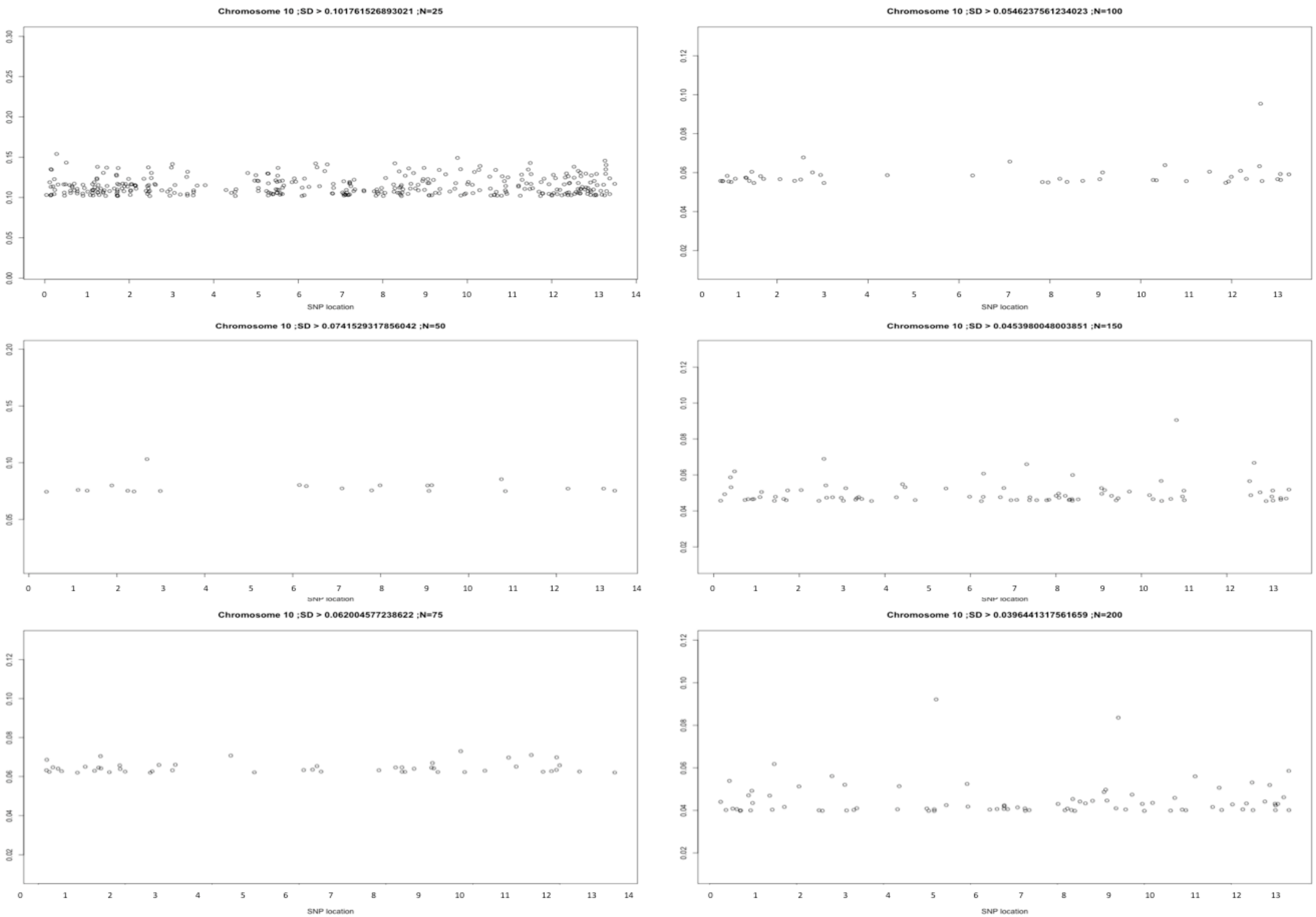


**Figure 4-10. Dot plots of outliers and their locations on chromosome 10 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

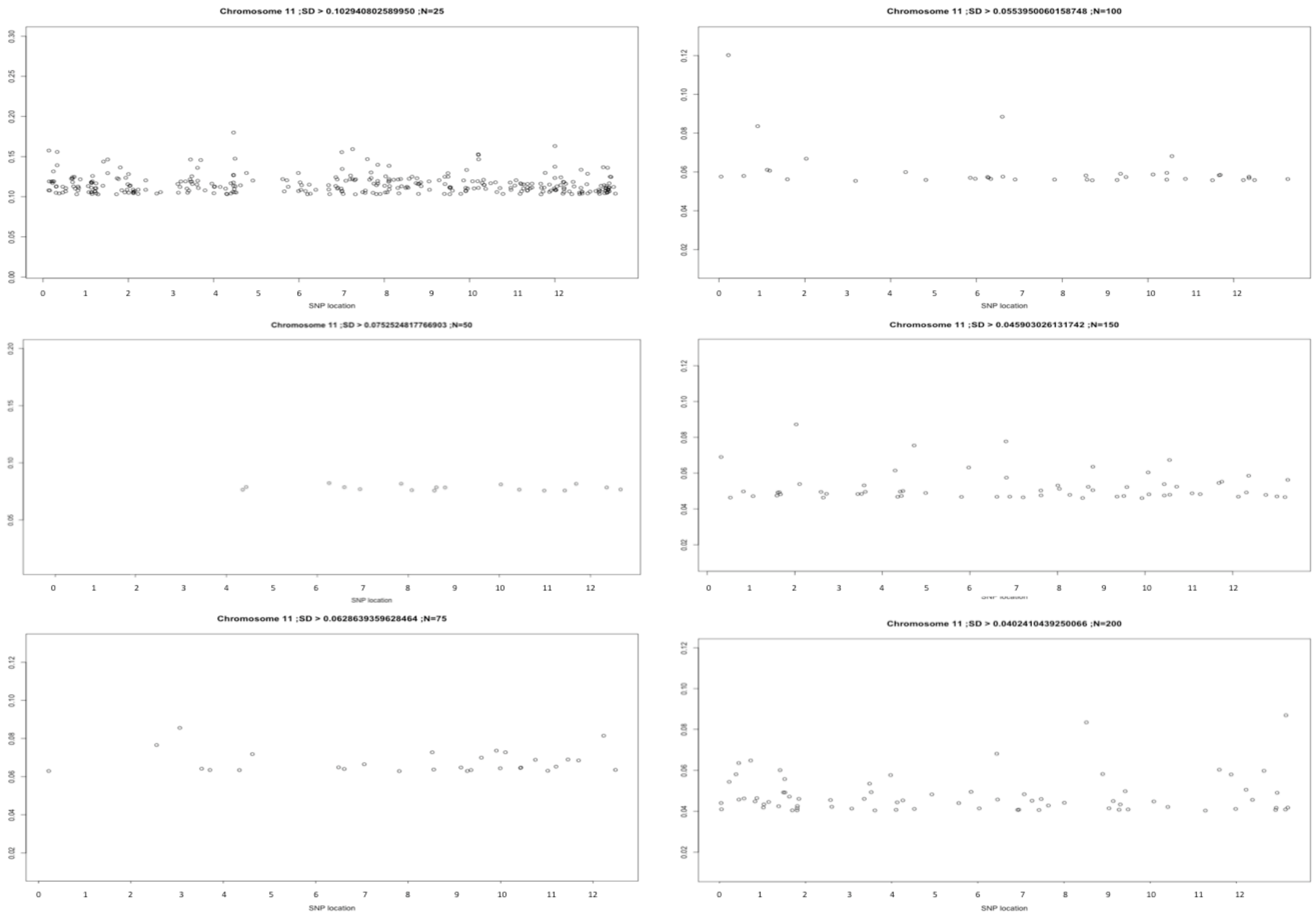


**Figure 4-11. Dot plots of outliers and their locations on chromosome 11 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

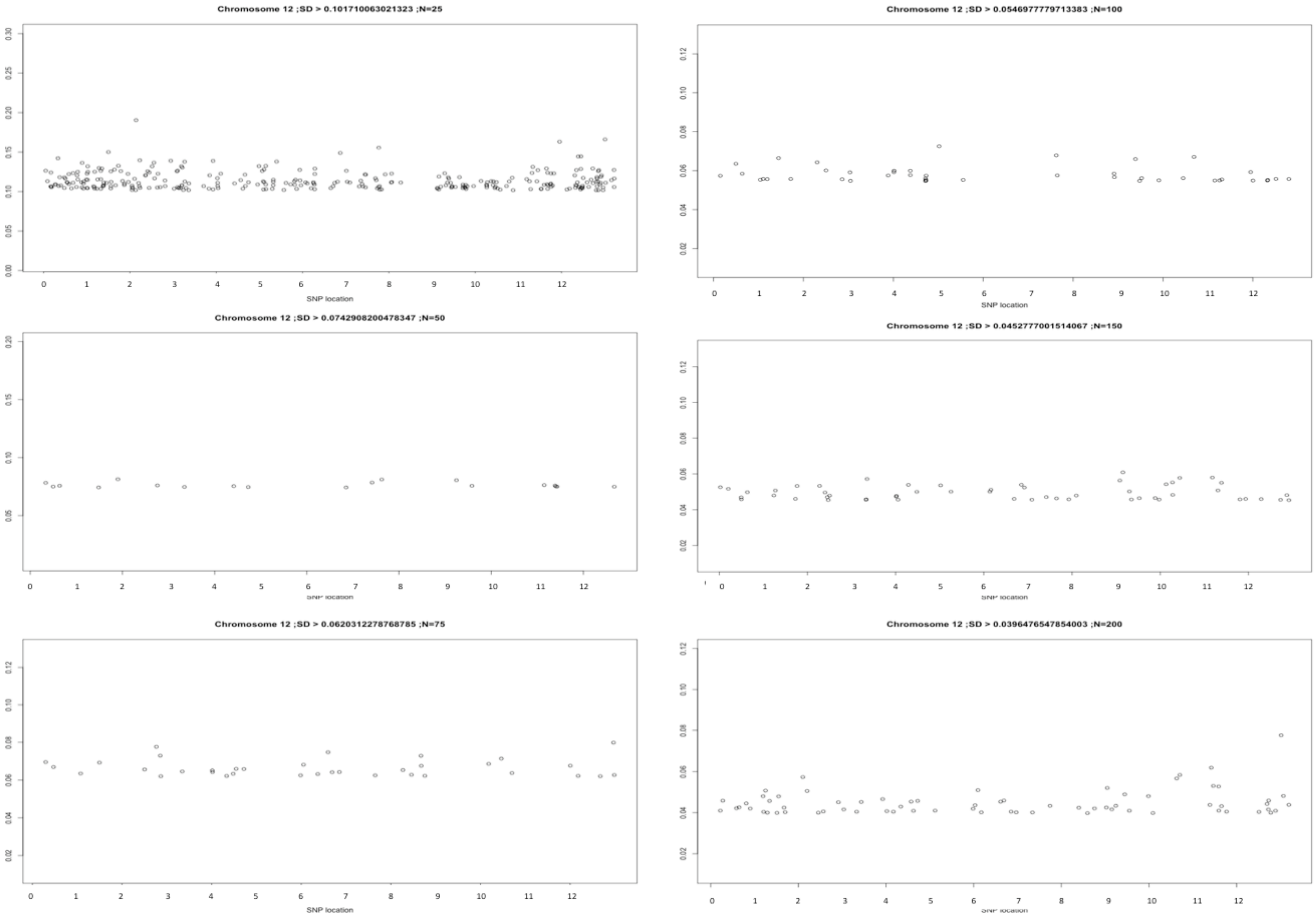


**Figure 4-12. Dot plots of outliers and their locations on chromosome 12 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

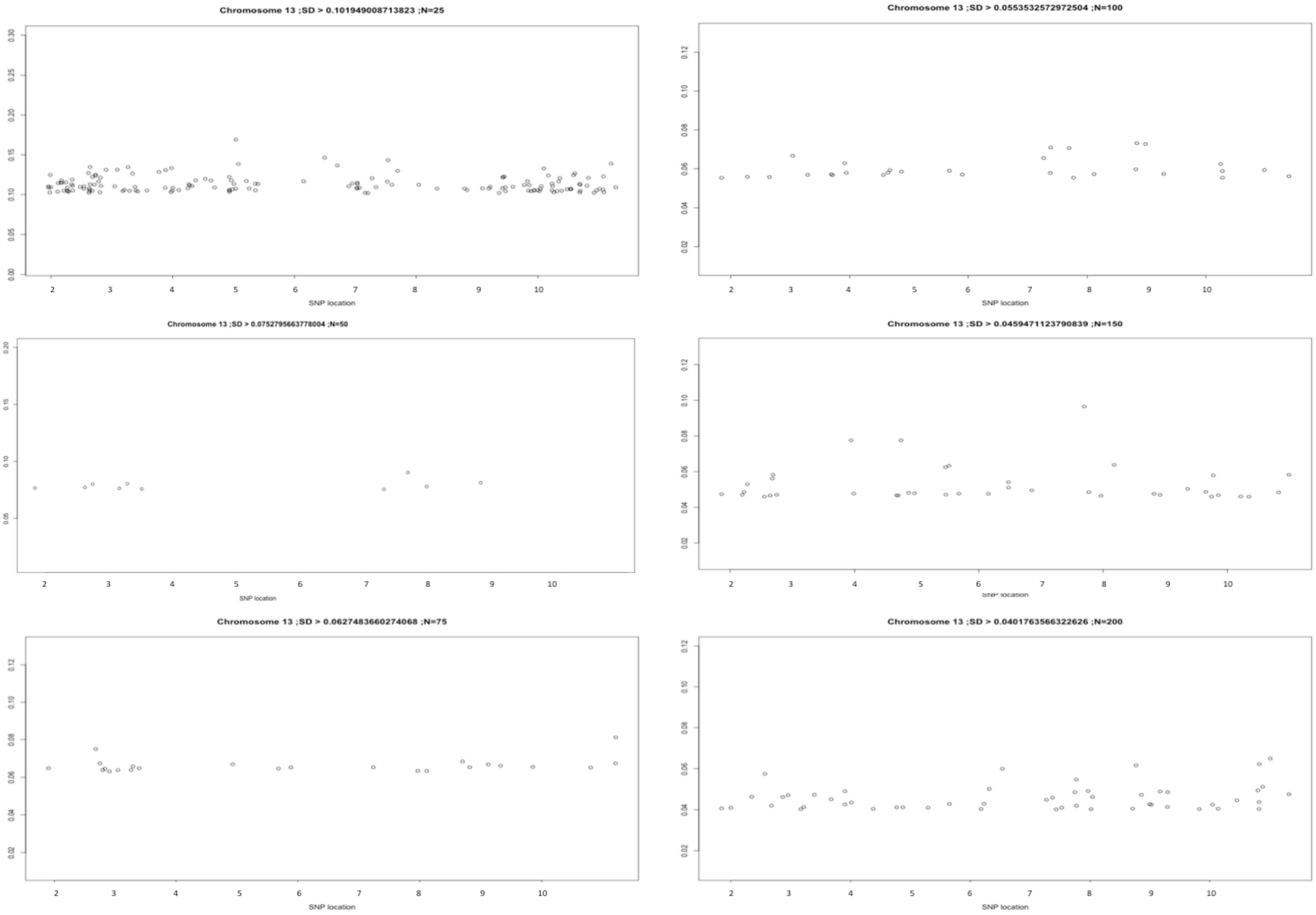


**Figure 4-13. Dot plots of outliers and their locations on chromosome 13 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

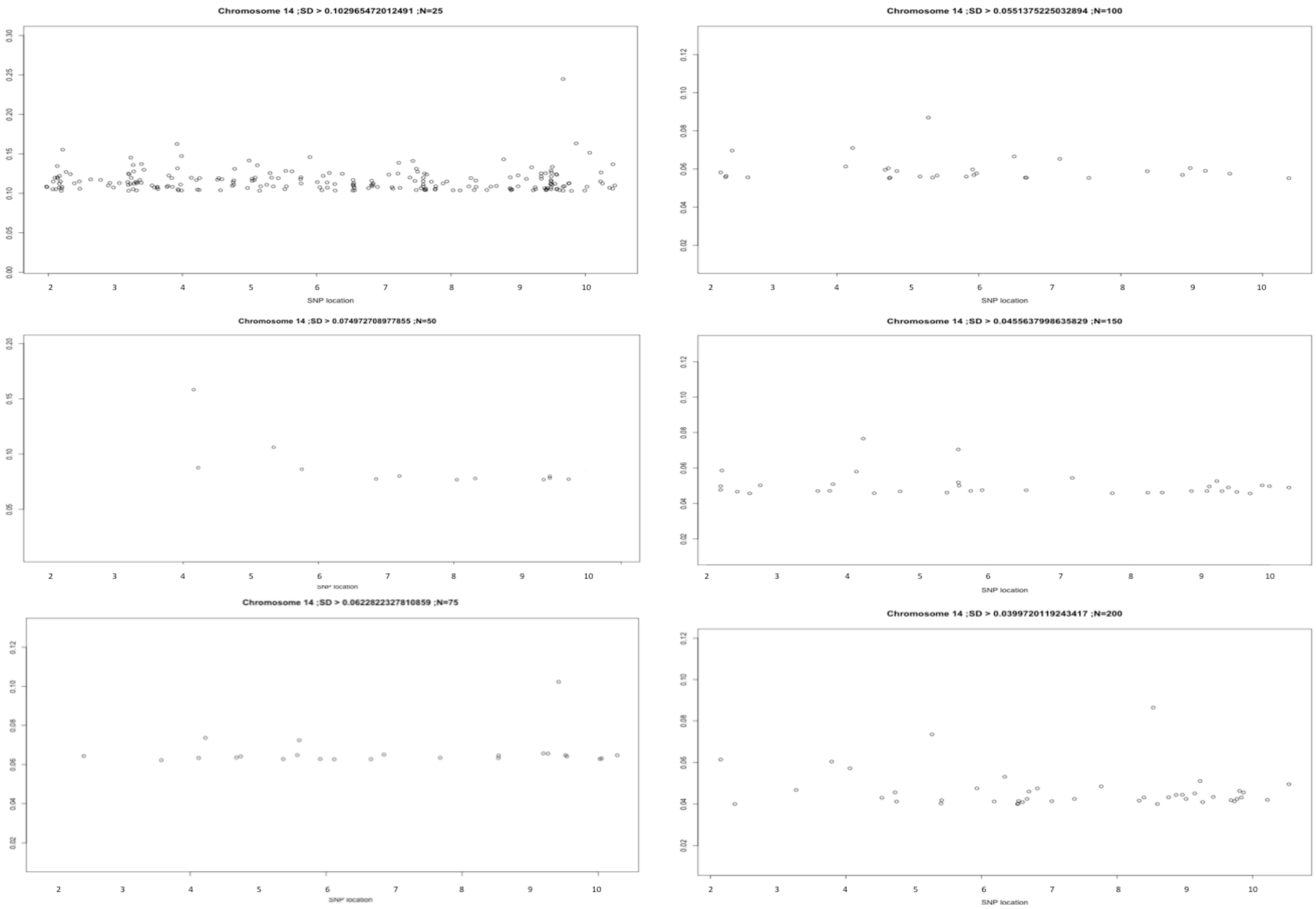


**Figure 4-14. Dot plots of outliers and their locations on chromosome 14 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

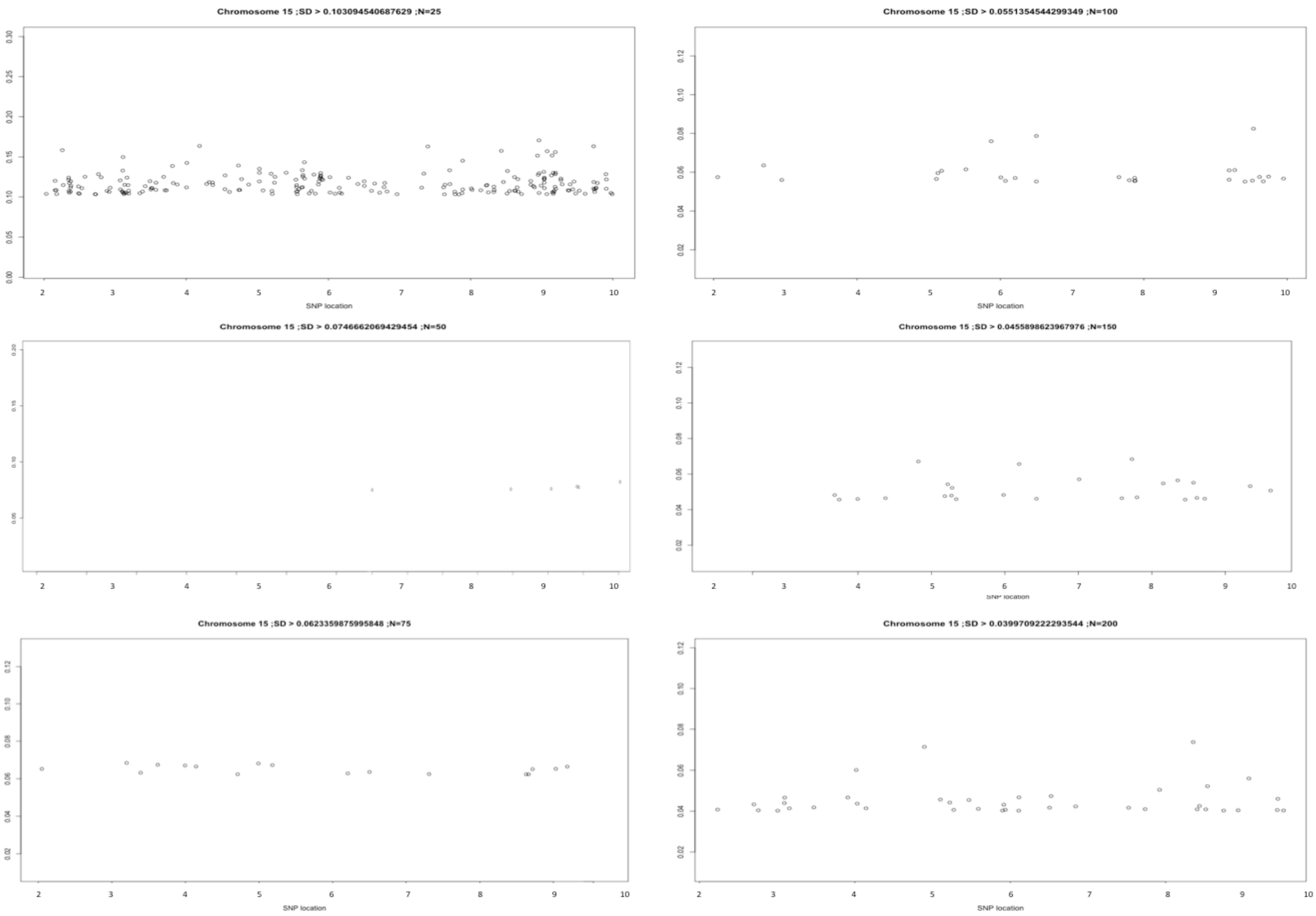


**Figure 4-15. Dot plots of outliers and their locations on chromosome 15 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

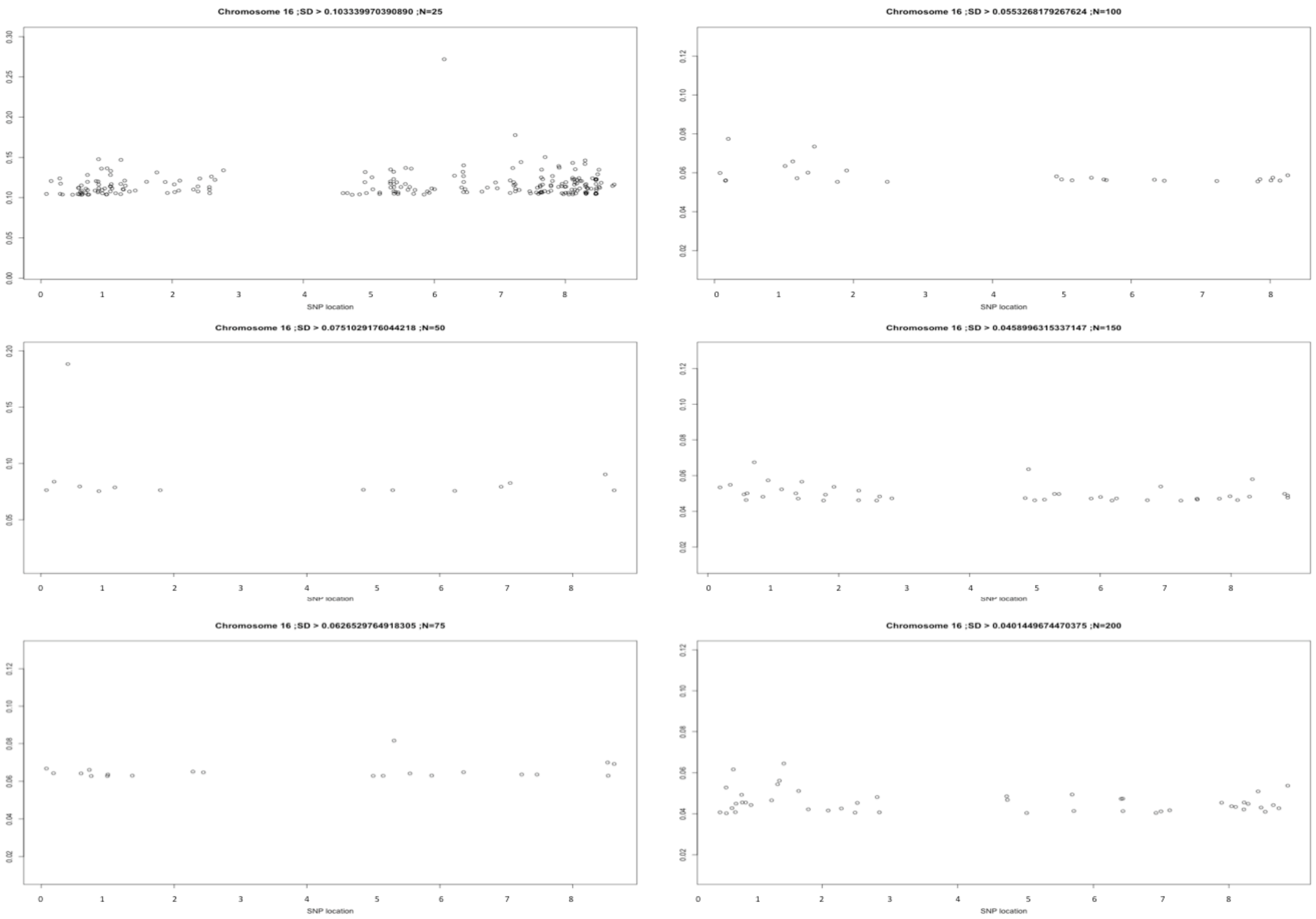


**Figure 4-16. Dot plots of outliers and their locations on chromosome 16 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

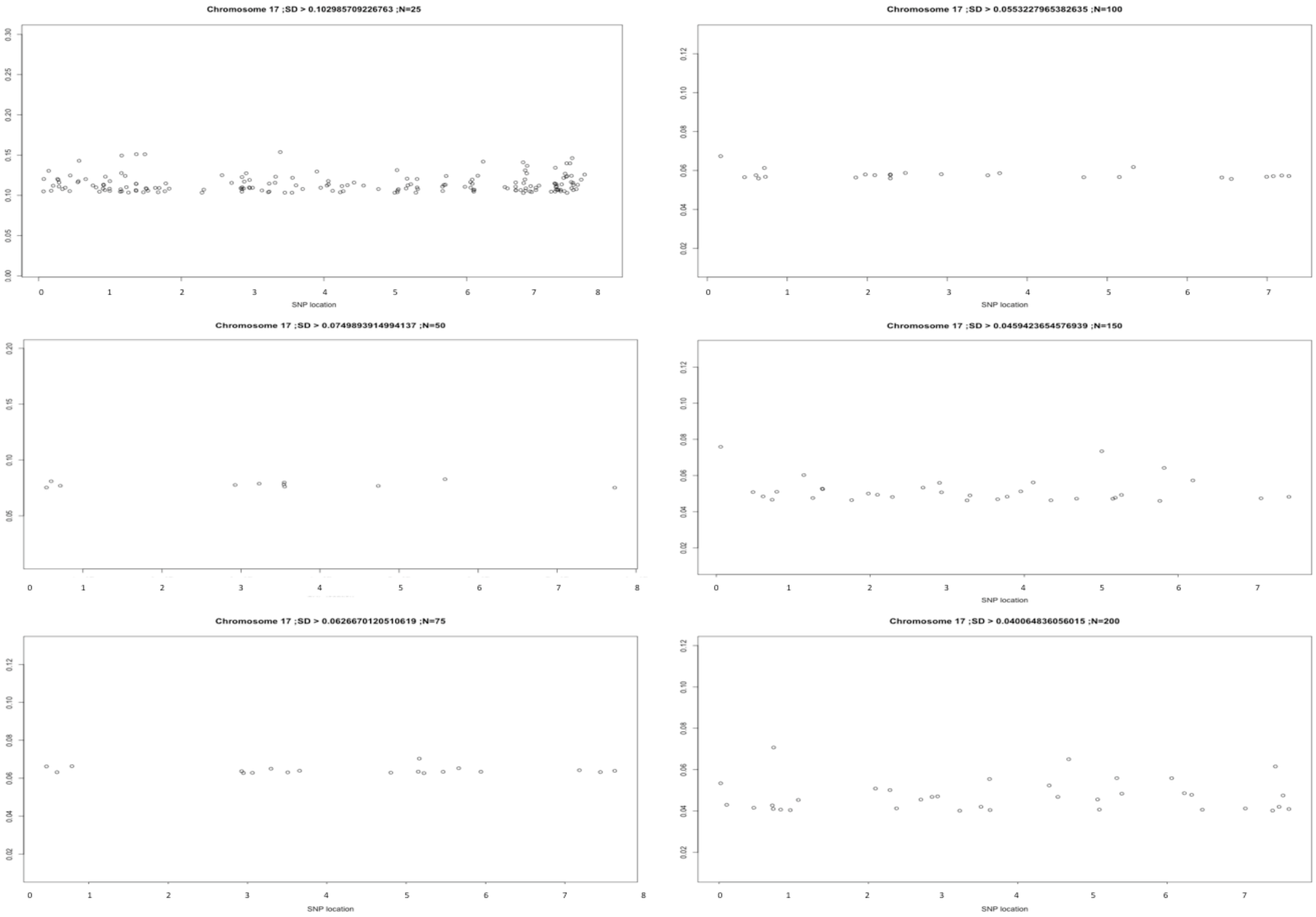


**Figure 4-17. Dot plots of outliers and their locations on chromosome 17 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

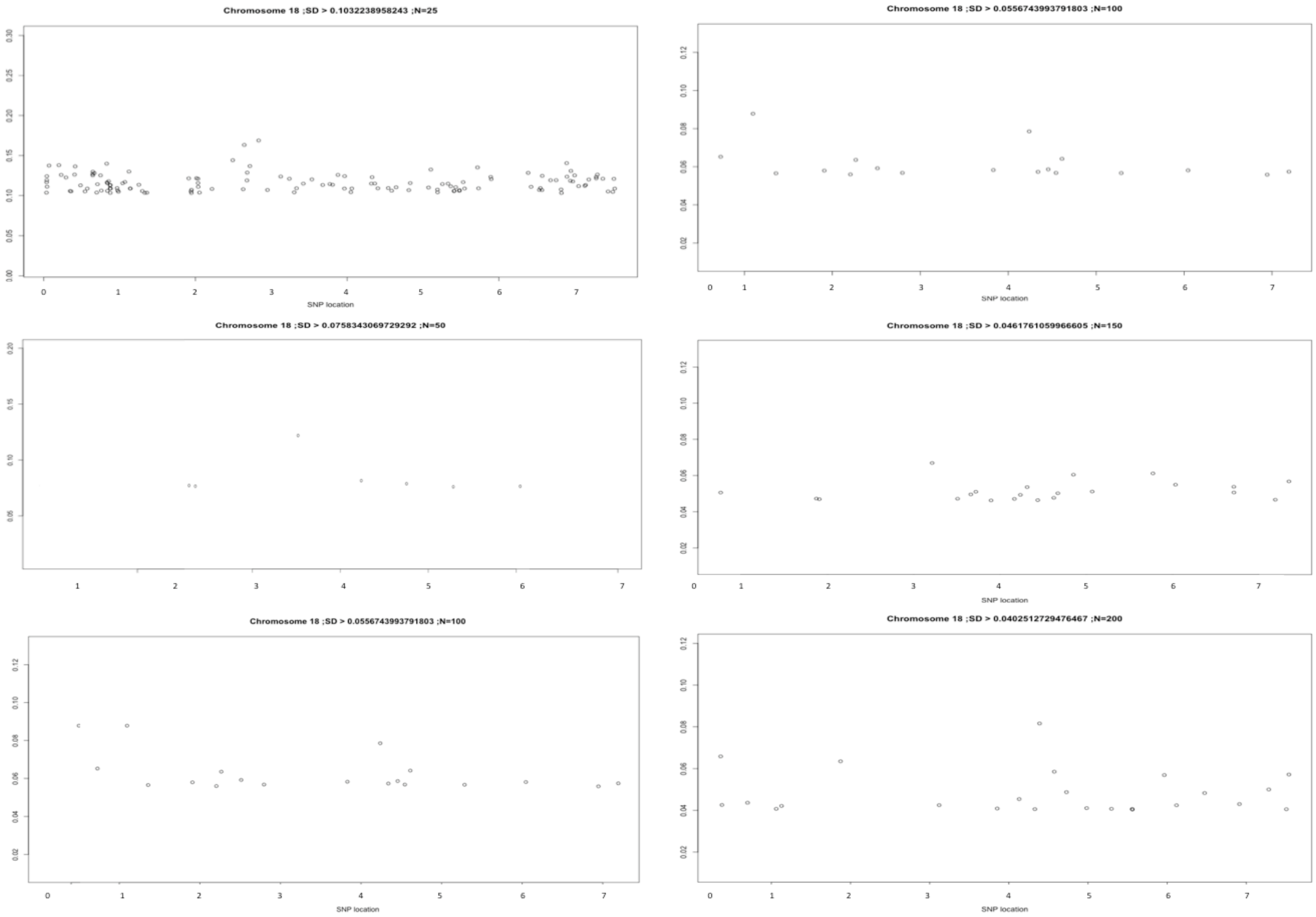


**Figure 4-18. Dot plots of outliers and their locations on chromosome 18 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

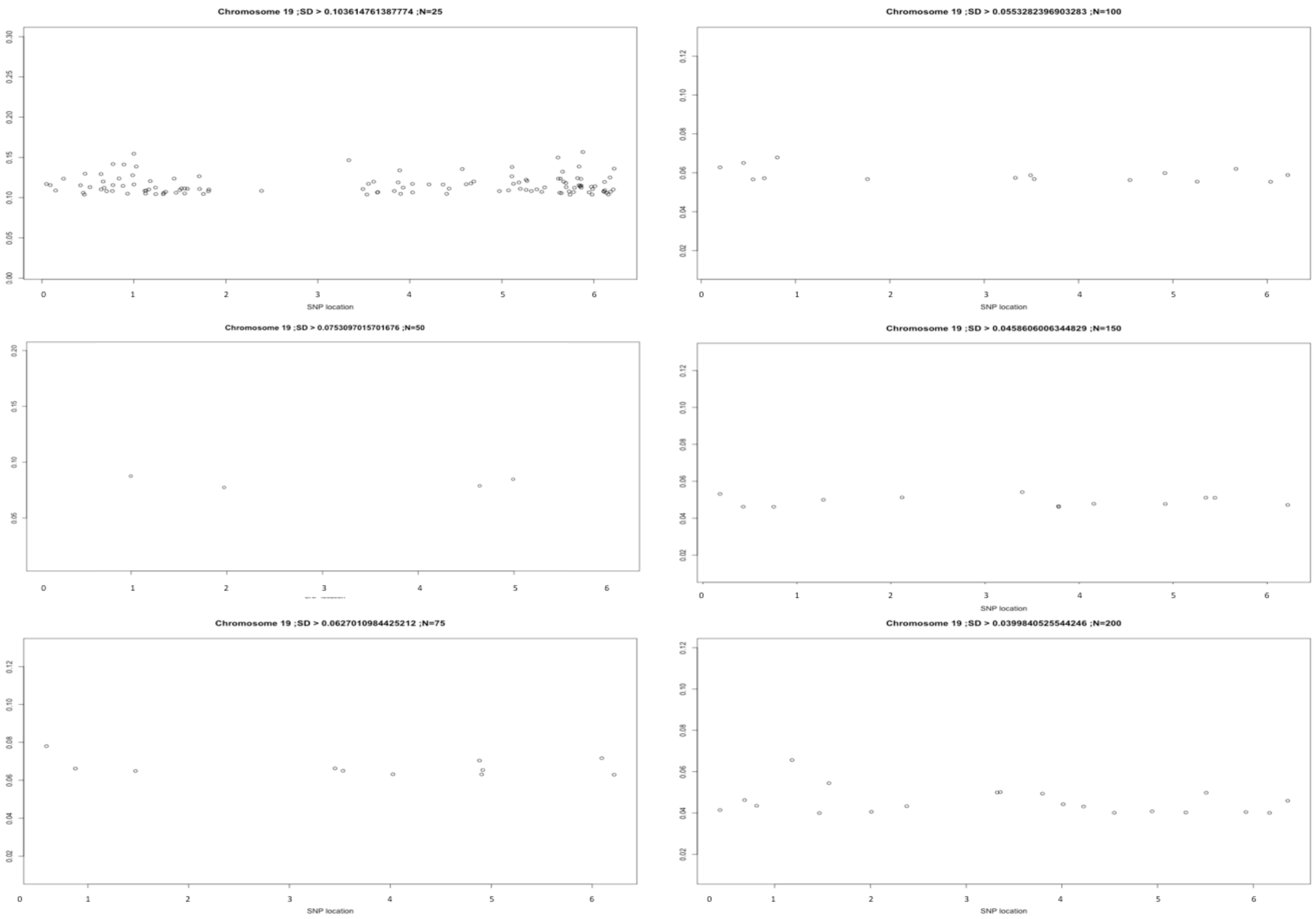


**Figure 4-19. Dot plots of outliers and their locations on chromosome 19 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

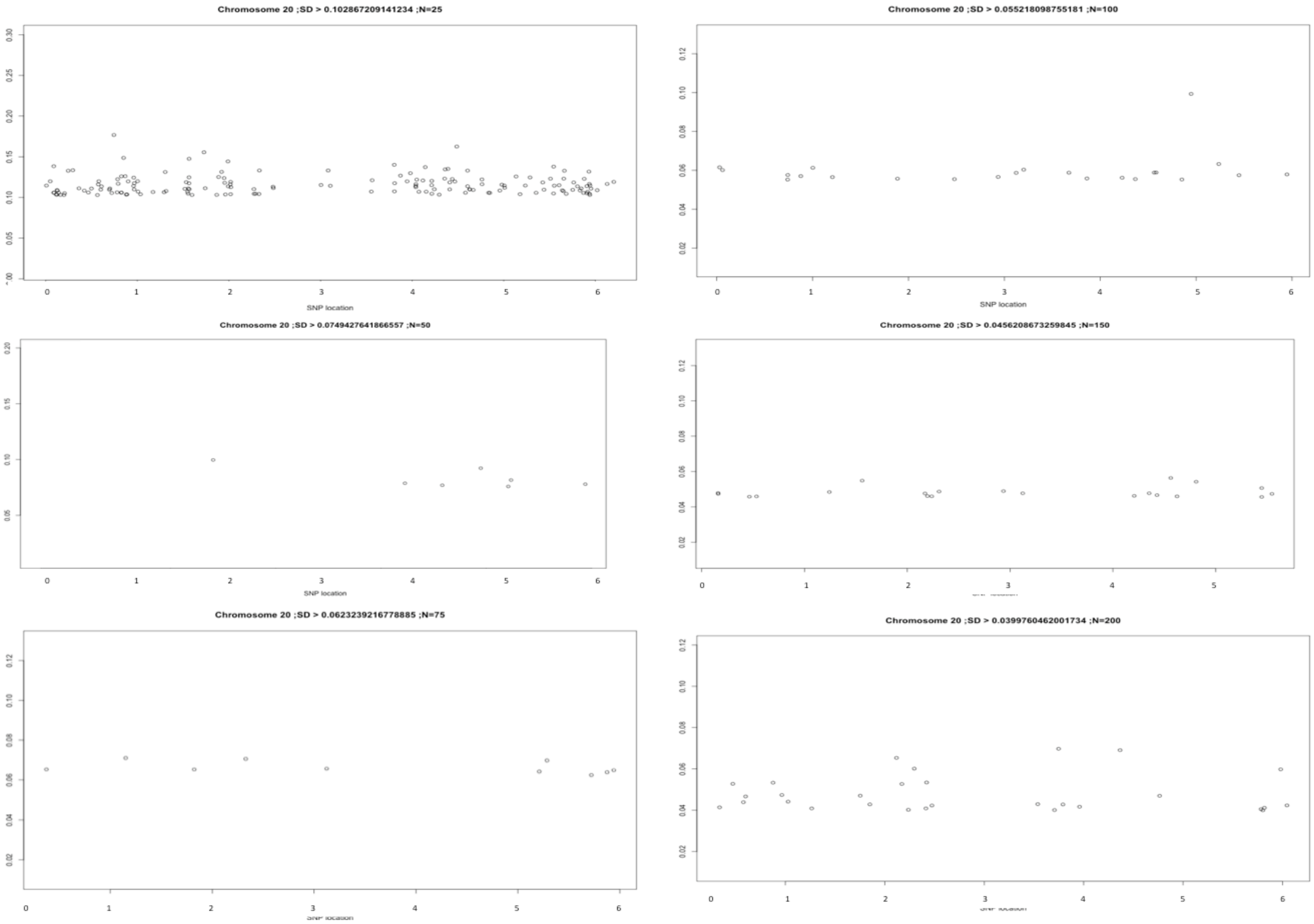


**Figure 4-20. Dot plots o outliers and their locations on chromosome 20 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

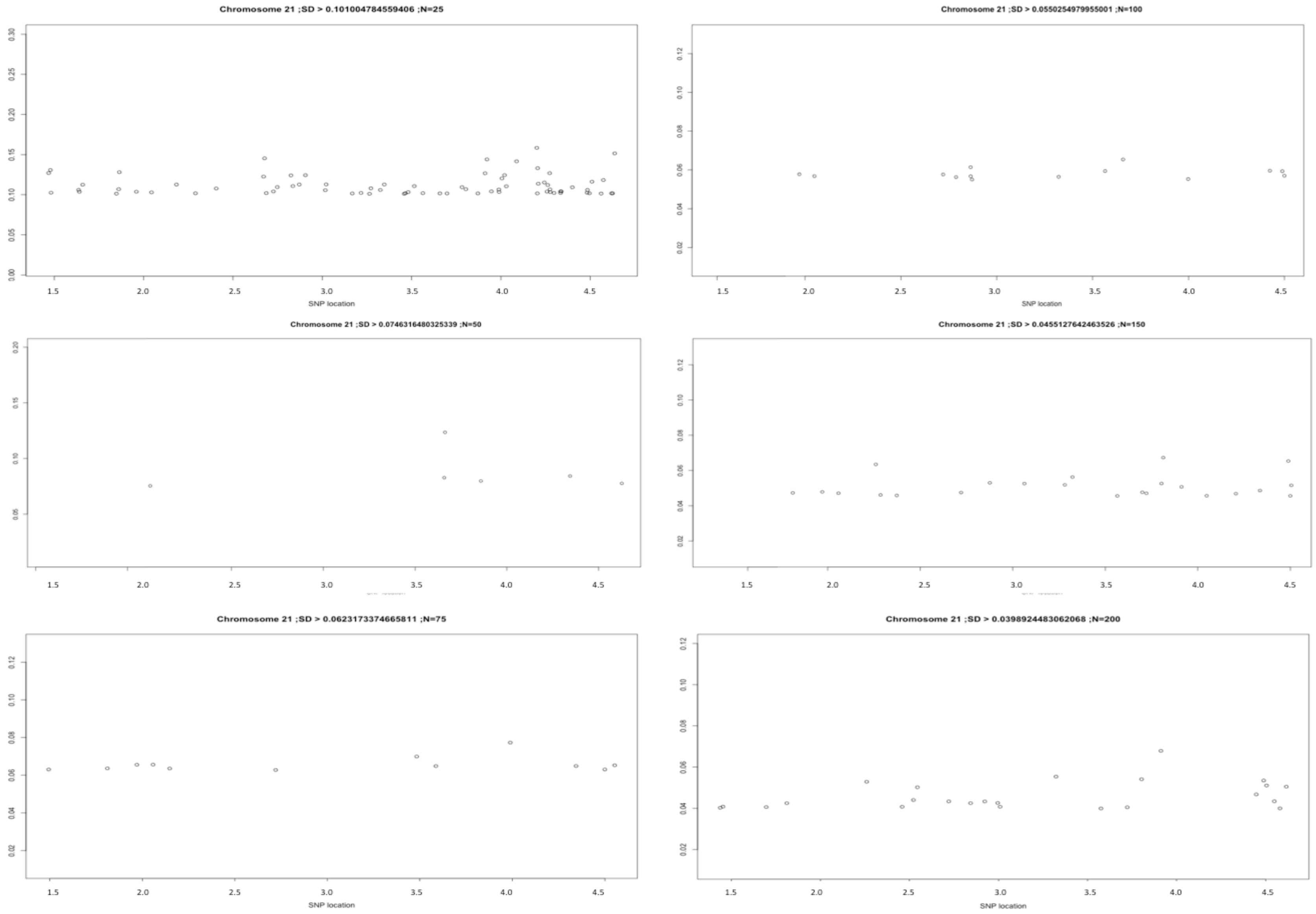


**Figure 4-21. Dot plots of outliers and their locations on chromosome 21 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

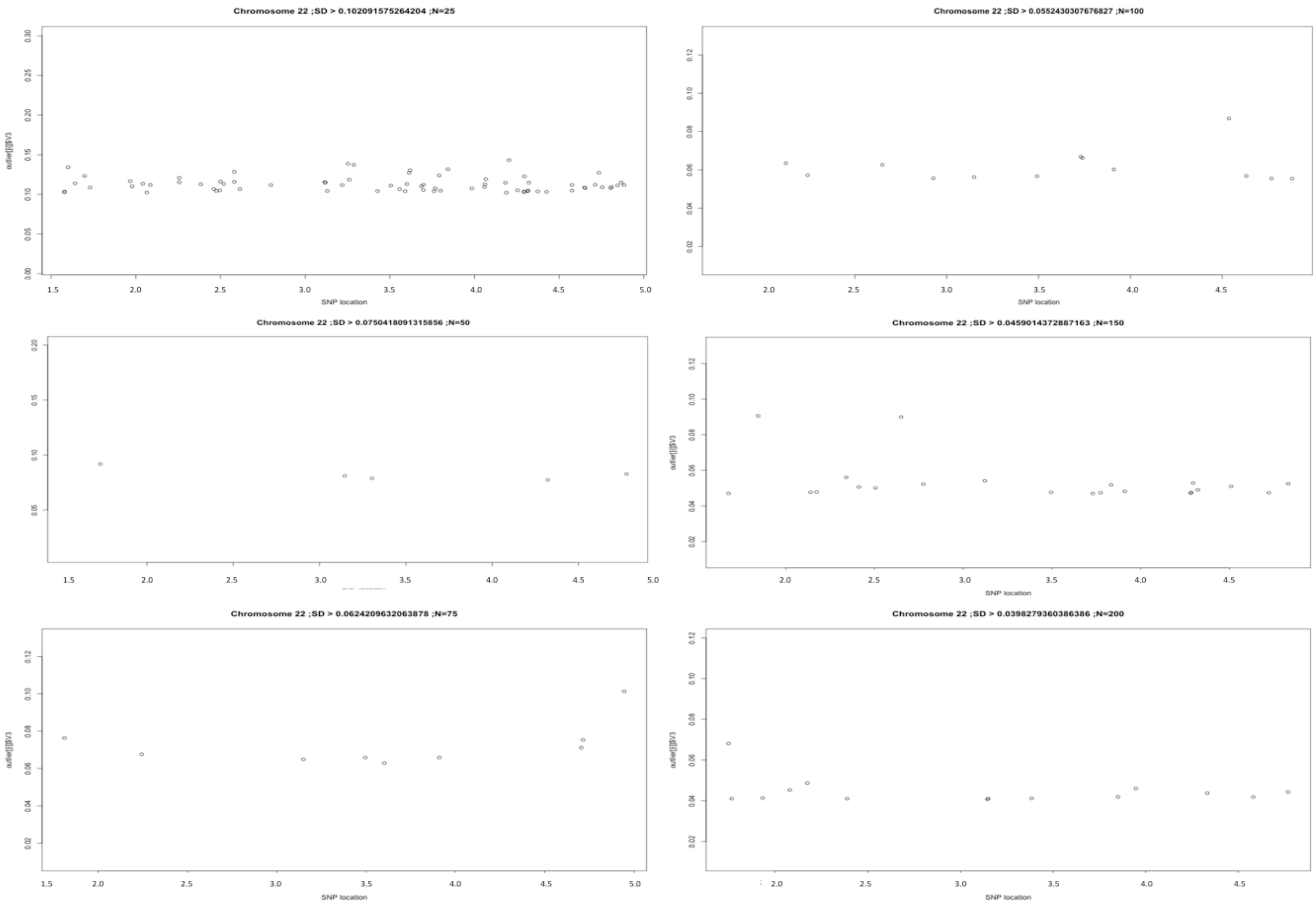


**Figure 4-22. Dot plots of the outliers and their locations on chromosome 22 (from left upper to right lower: n=25, 50, 75, 100, 150, and 200)**

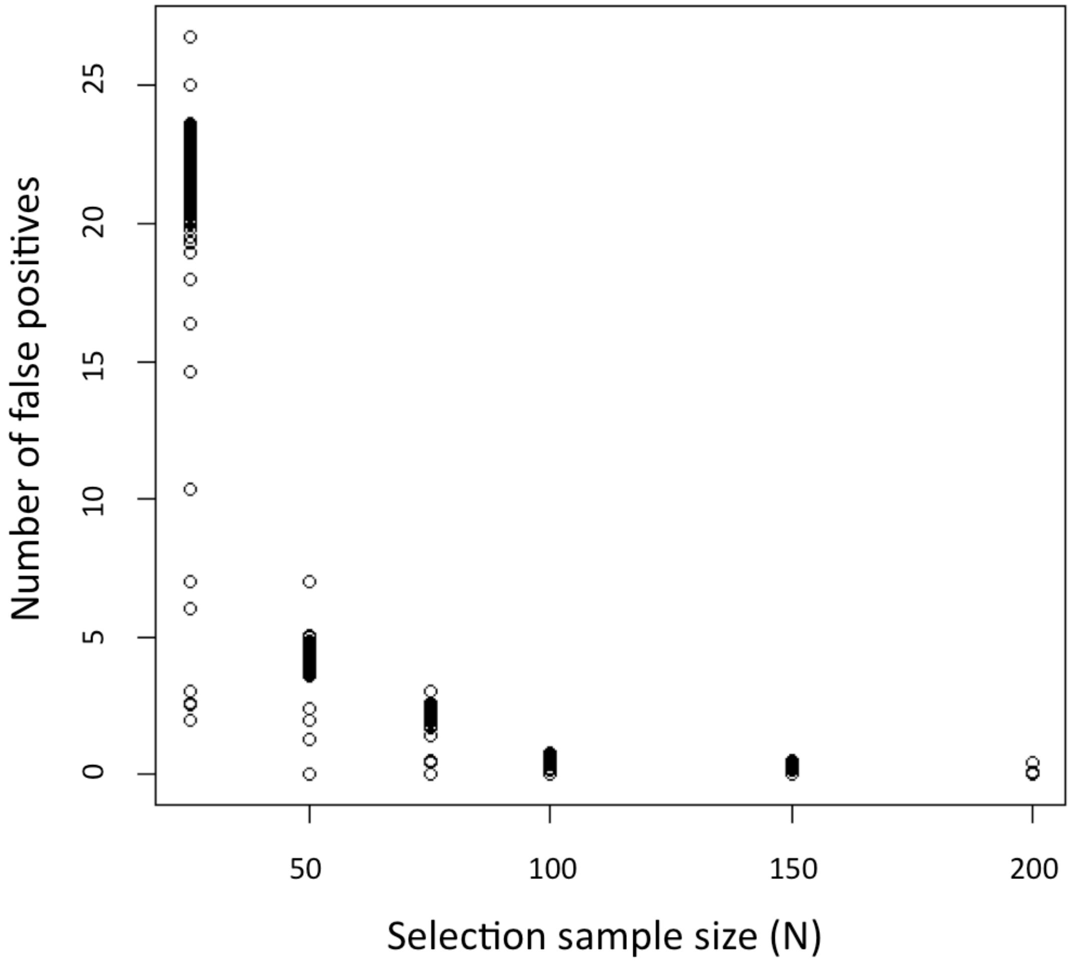


**Figure 5-1. Scatter plot of sample size and average number of false positives**

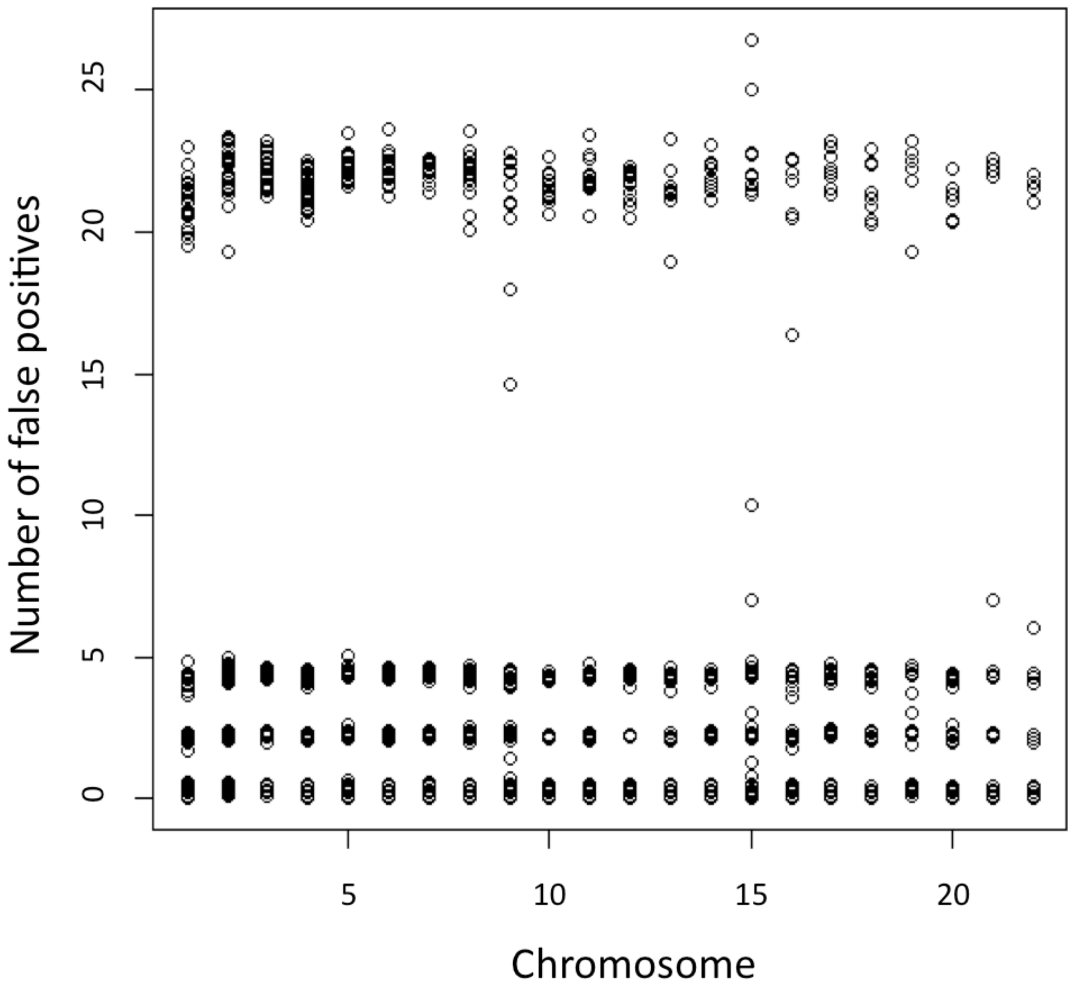


**Figure 5-2. Scatter plot of chromosome and average number of false positives**

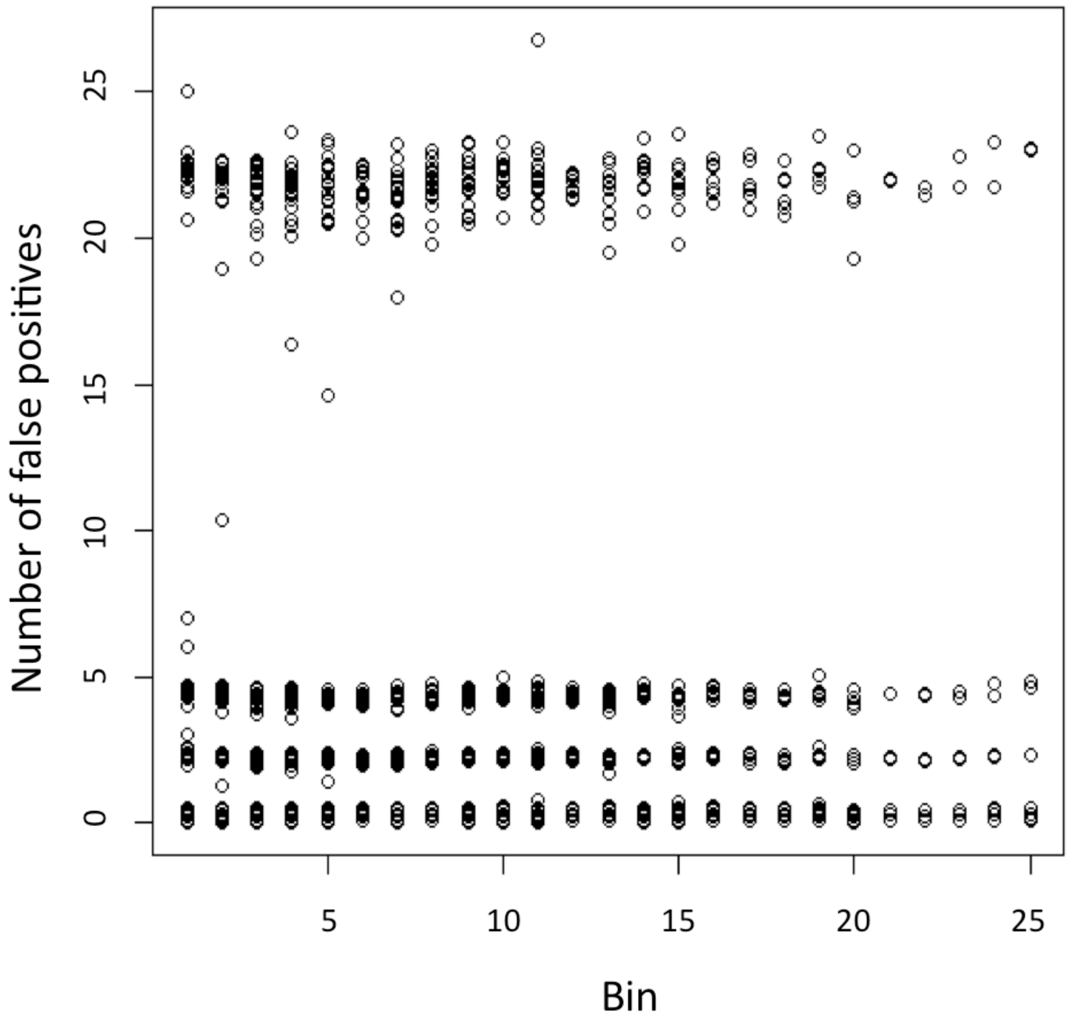


**Figure 5-3. Scatter plot of bin and average number of false positives**

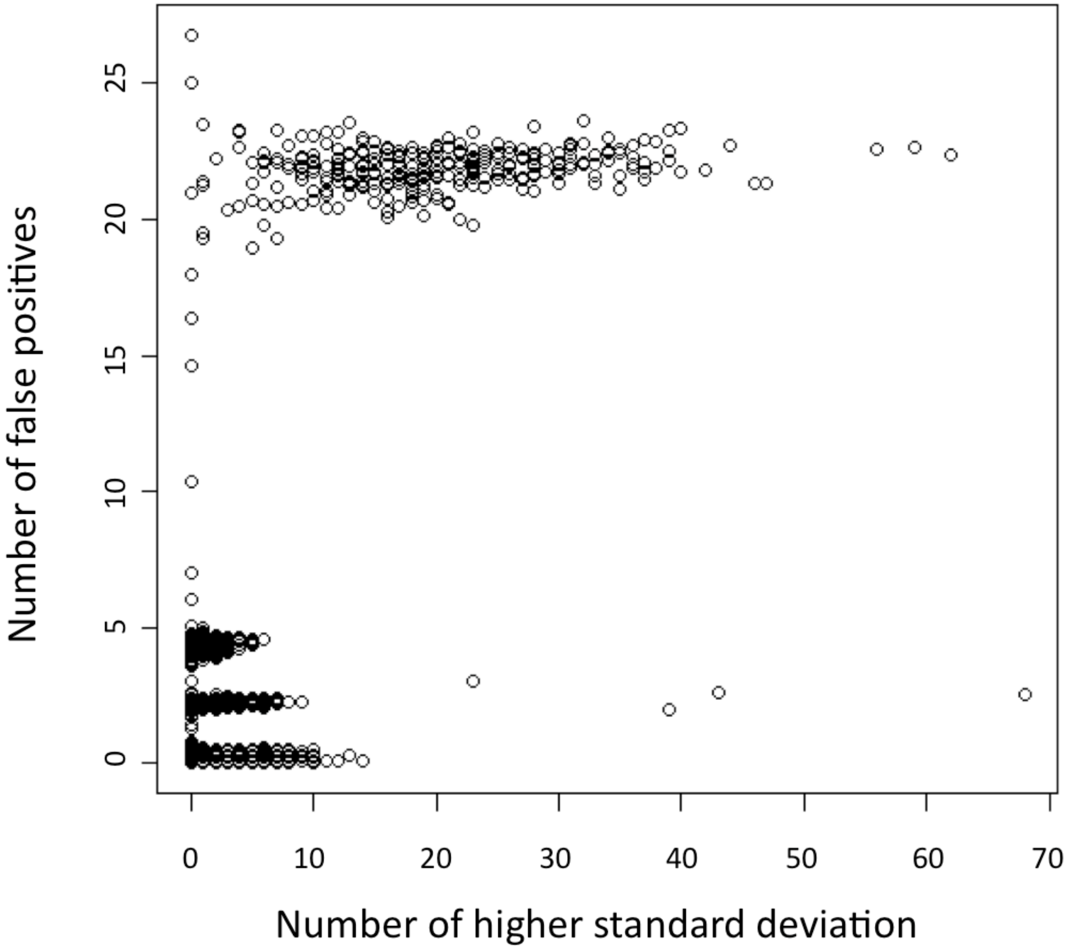


**Figure 5-4. Scatter plot of number of higher standard deviation and average number of false positives**

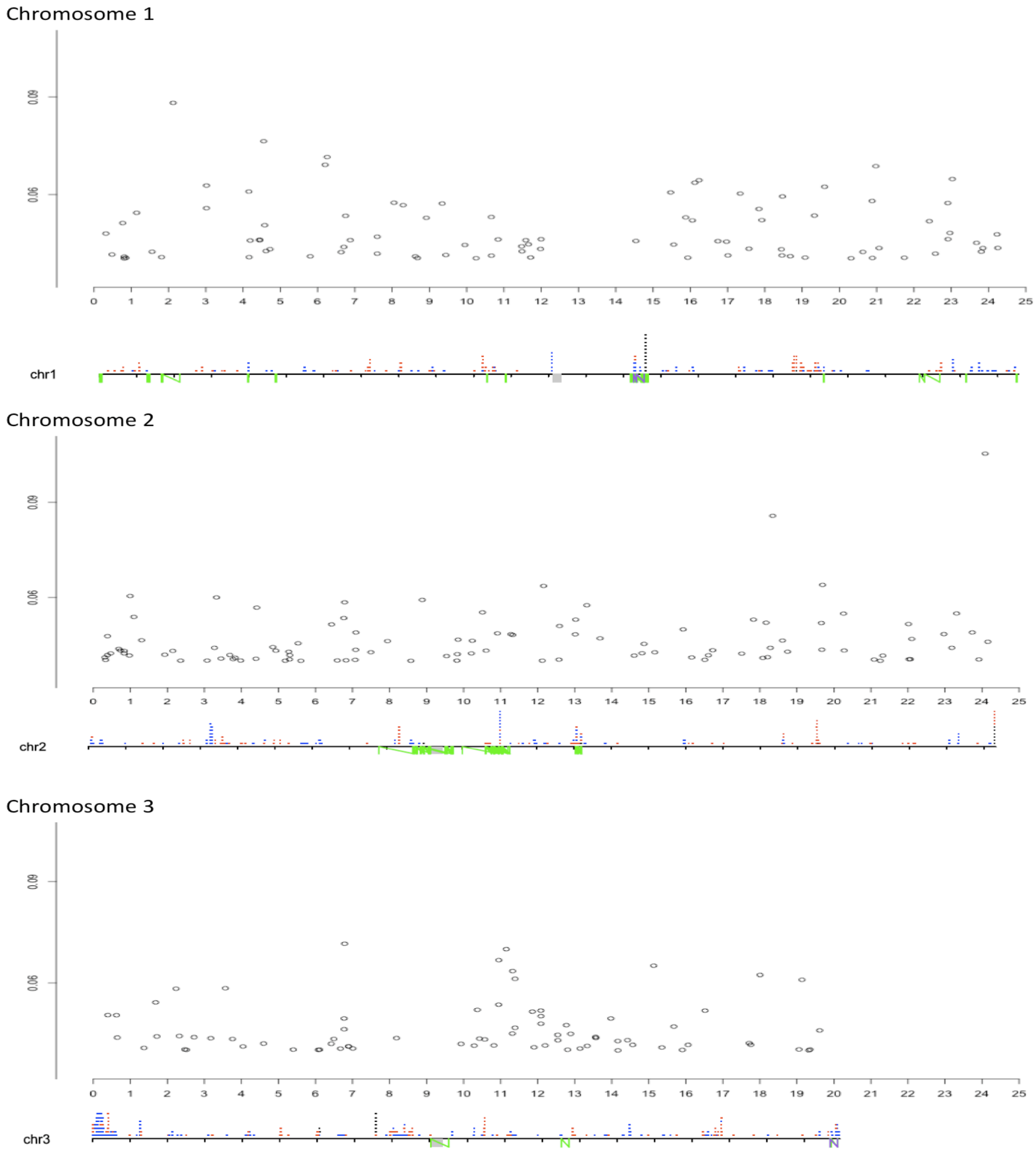
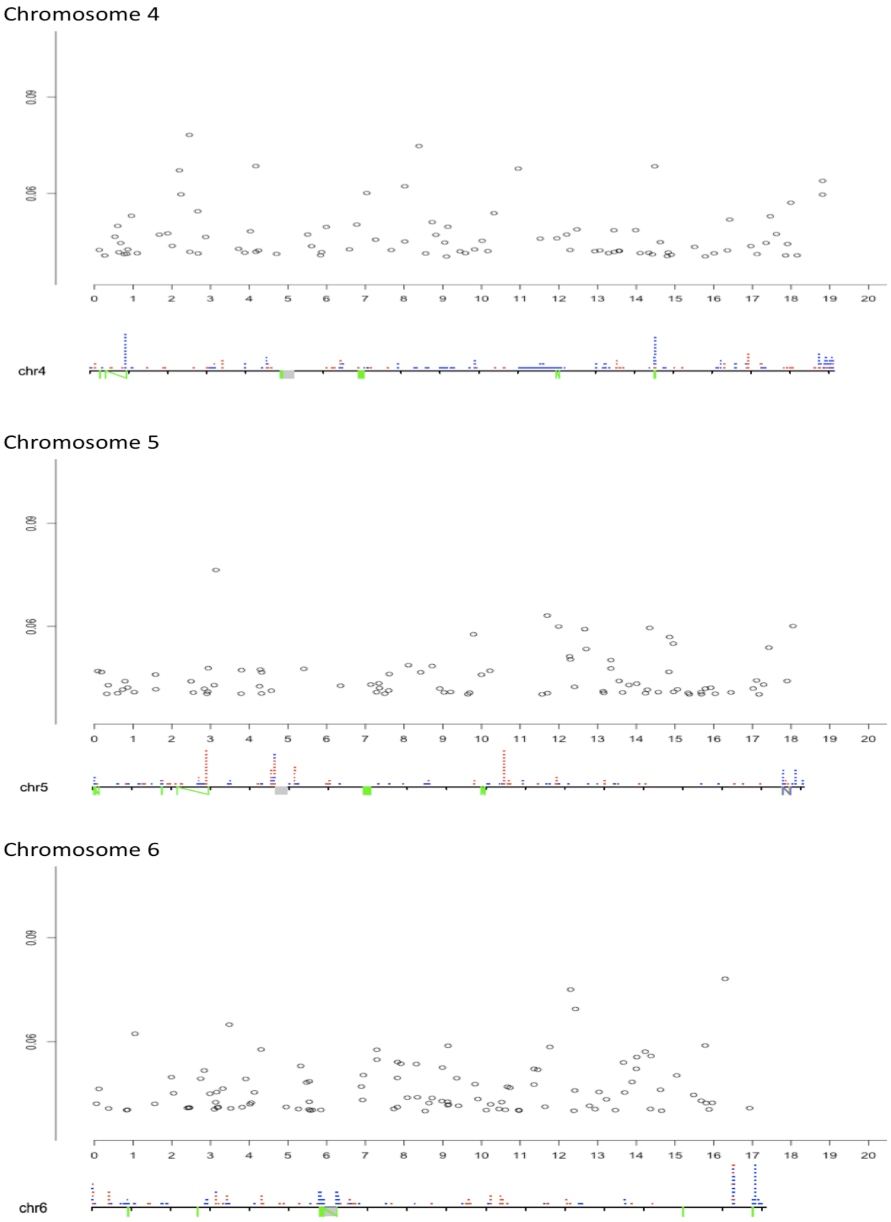


**Figure 6-1. Comparison between Itsara et al's autosomal landscape of large CNV plots (below) and our outlier plots (above) for chromosomes 1 to 3. Adapted with permission from figure 2 of Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, Mefford H, Ying P, Nickerson DA, Eichler EE (2009) Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 84:148-161.**

* On Itsara et al plots, blue dots indicates the duplications, red dots indicates deletions, black dots indicates homozygous deletions, gray dots indicates centromeres, green lines indicates predicted rearrangement hotspots, and purple indicates hotspots associated with diseases.



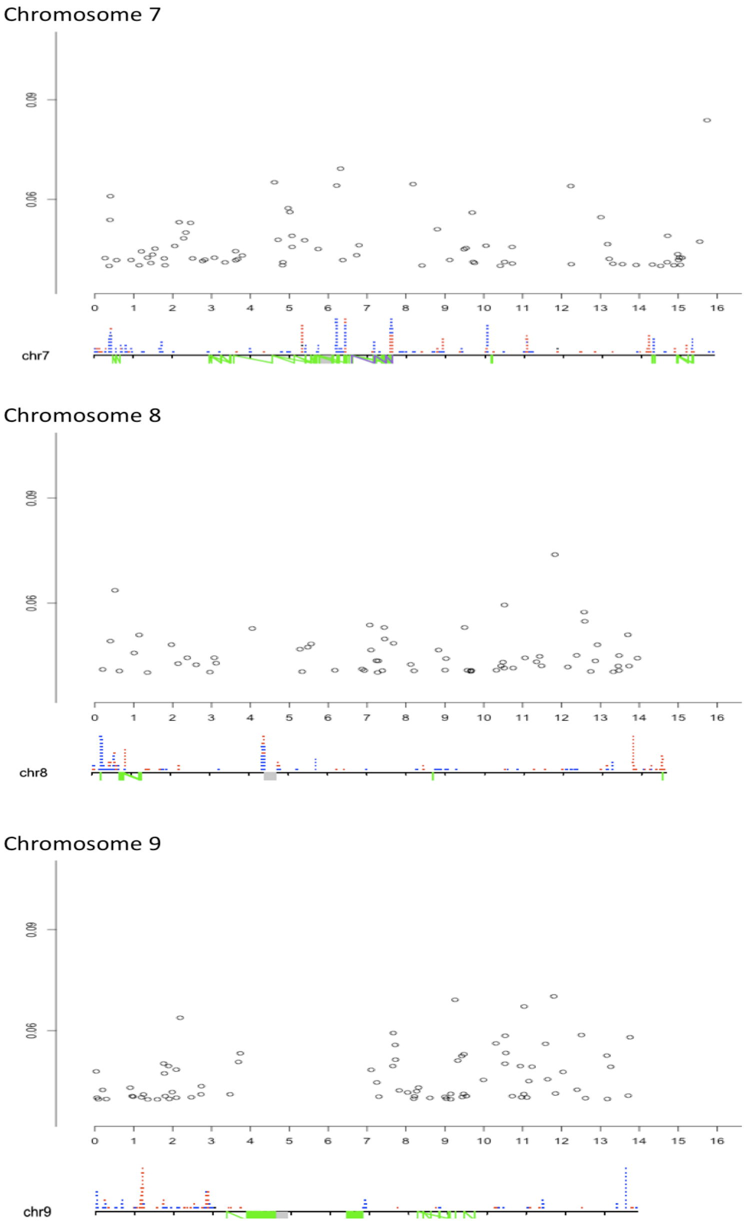
Chromosome 4

Chromosome 5

Chromosome 6

**Figure 6-2. Comparison between Itsara et al's autosomal landscape of large CNV plots (below) and our outlier plots (above) for chromosomes 4 to 6. Adapted with permission from figure 2 of Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, Mefford H, Ying P, Nickerson DA, Eichler EE (2009) Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 84:148-161.**

* On Itsara et al plots, blue dots indicates the duplications, red dots indicates deletions, black dots indicates homozygous deletions, gray dots indicates centromeres, green lines indicates predicted rearrangement hotspots, and purple indicates hotspots associated with diseases.



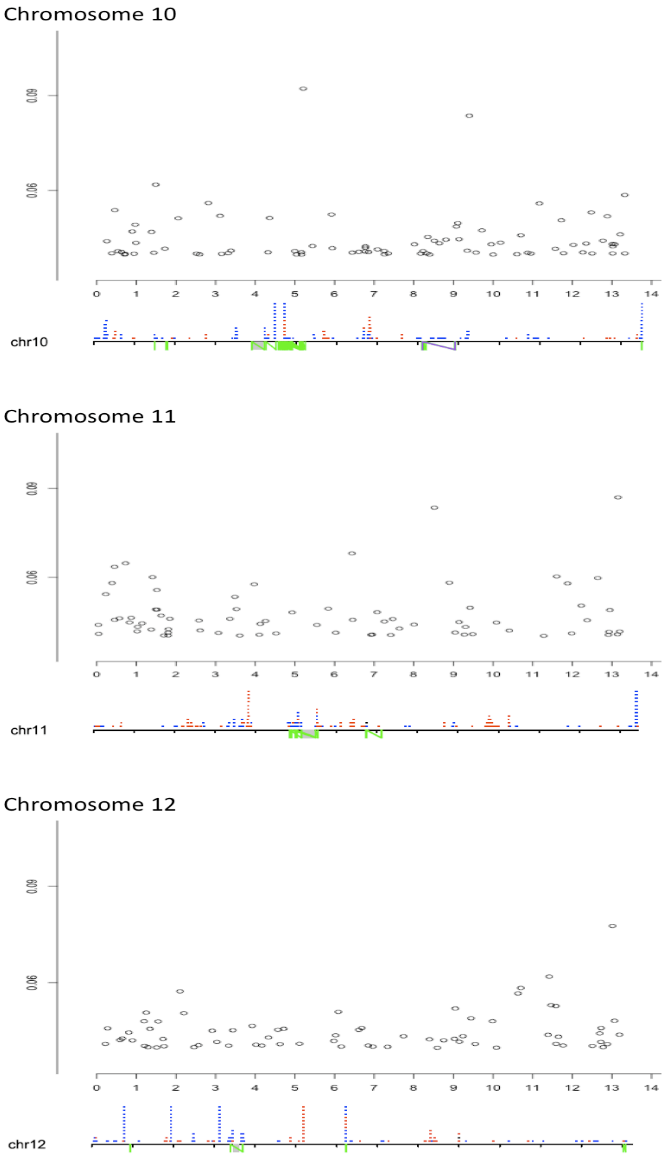
Chromosome 7

Chromosome 8

Chromosome 9

**Figure 6-3. Comparison between Itsara et al's autosomal landscape of large CNV plots (below) and our outlier plots (above) for chromosomes 7 to 9. Adapted with permission from figure 2 of Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, Mefford H, Ying P, Nickerson DA, Eichler EE (2009) Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 84:148-161.**

* On Itsara et al plots, blue dots indicates the duplications, red dots indicates deletions, black dots indicates homozygous deletions, gray dots indicates centromeres, green lines indicates predicted rearrangement hotspots, and purple indicates hotspots associated with diseases.


Chromosome 10

Chromosome 11

Chromosome 12

**Figure 6-4. Comparison between Itsara et al's autosomal landscape of large CNV plots (below) and our outlier plots (above) for chromosomes 10 to 12. Adapted with permission from figure 2 of Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, Mefford H, Ying P, Nickerson DA, Eichler EE (2009) Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 84:148-161.**

* On Itsara et al plots, blue dots indicates the duplications, red dots indicates deletions, black dots indicates homozygous deletions, gray dots indicates centromeres, green lines indicates predicted rearrangement hotspots, and purple indicates hotspots associated with diseases.

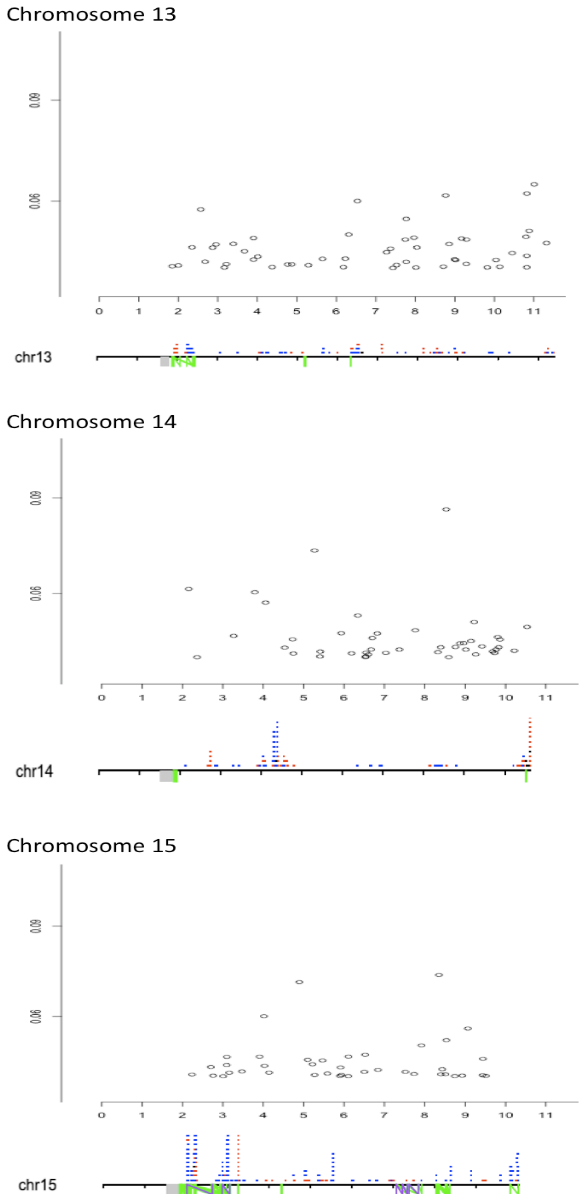
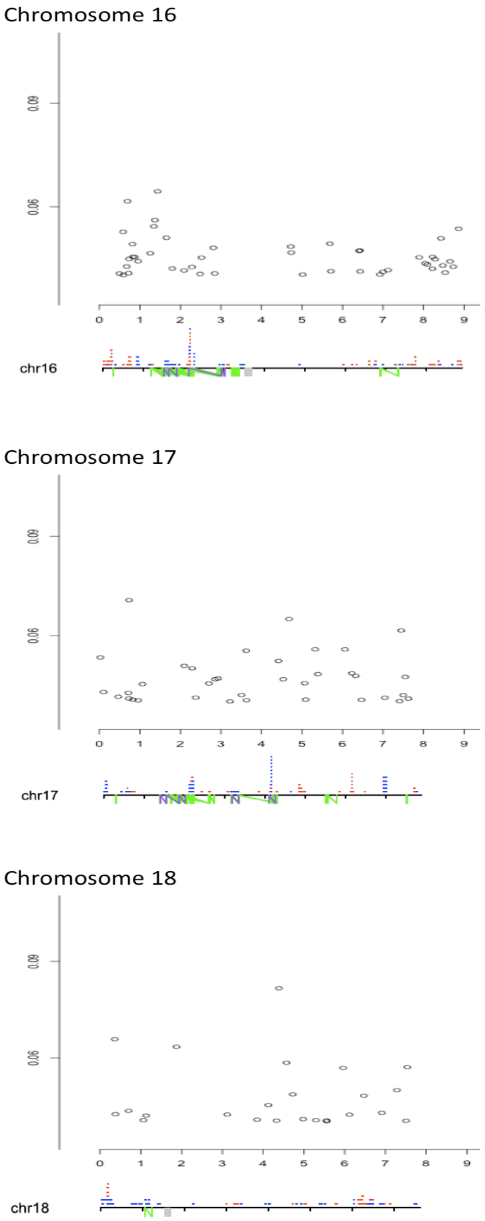


**Figure 6-5. Comparison between Itsara et al's autosomal landscape of large CNV plots (below) and our outlier plots (above) for chromosomes 13 to 15. Adapted with permission from figure 2 of Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, Mefford H, Ying P, Nickerson DA, Eichler EE (2009) Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 84:148-161.**

* On Itsara et al plots, blue dots indicates the duplications, red dots indicates deletions, black dots indicates homozygous deletions, gray dots indicates centromeres, green lines indicates predicted rearrangement hotspots, and purple indicates hotspots associated with diseases.


Chromosome 16

Chromosome 17

Chromosome 18

**Figure 6-6. Comparison between Itsara et al's autosomal landscape of large CNV plots (below) and our outlier plots (above) for chromosomes 16 to 18. Adapted with permission from figure 2 of Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, Mefford H, Ying P, Nickerson DA, Eichler EE (2009) Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 84:148-161.**

* On Itsara et al plots, blue dots indicates the duplications, red dots indicates deletions, black dots indicates homozygous deletions, gray dots indicates centromeres, green lines indicates predicted rearrangement hotspots, and purple indicates hotspots associated with diseases.

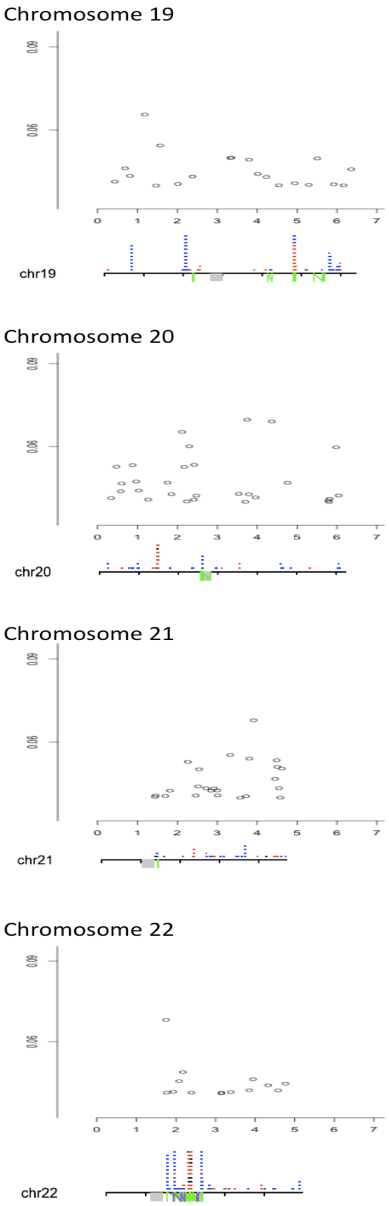


**Figure 6-7. Comparison between Itsara et al‘s autosomal landscape of large CNV plots (below) and our outlier plots (above) for chromosomes 19 to 22. Adapted with permission from figure 2 of Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, Mefford H, Ying P, Nickerson DA, Eichler EE (2009) Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 84:148-161.**

* On Itsara et al plots, blue dots indicates the duplications, red dots indicates deletions, black dots indicates homozygous deletions, gray dots indicates centromeres, green lines indicates predicted rearrangement hotspots, and purple indicates hotspots associated with diseases.

# APPENDIX

## UNIX AND R CODE

| Step | Unix or R code | Notes |
|---|---|---|
| 1 | chmod +x script<br>./script 500 10078 200 | *500: replicates; 10078: overall sample size; 200: selection subjects |
| 2 | **% script**<br>#!/bin/tcsh -xv<br>./callselect.sh $1 $2 $3<br>./PLINKidlist.sh $1 | |

| | | |
|---|---|---|
| | ./freq.sh $1<br>./association.sh<br>./location.sh<br>Exit | |
| 3 | **% callselect.sh**<br>#!/bin/tcsh -vx<br>echo "$2, $3" > num.tmp<br>@ i = 1<br>while ($i <= $1)<br>R CMD BATCH select.R<br>cp c.out idlist$i.txt<br>@ i++<br>end<br>exit | * Create lists of selected subjects, end up with 500 id lists. |
| 4 | **% select.R**<br>tmp <- read.csv("num.tmp",header=F) | |

| | | |
|---|---|---|
| | a <- read.table("list.txt",header=F)<br>id <-sample(c(1:nrow(a)),tmp[1,2],replace=FALSE,rep(0.1,nrow(a)))<br>b <- rep(0,nrow(a))<br>for (i in 1:tmp[1,2])<br>{<br>b[id[i]] <- 1<br>}<br>data <- cbind(b,a)<br>c <- data[which(data[,1]=="1"),2:3]<br>write.table(c,"c.out",sep="\t",row.names=FALSE,col.names=FALSE<br>, quote=FALSE) | * ‘list.txt’ contains family IDs and subject IDs<br>* Subjects were randomly selected from the whole study populaton. |
| 5 | **% PLINKidlist.sh**<br>#!/bin/tcsh -vx<br>@ i = 1<br>while ($i <= $1)<br>plink --bfile final.merged --keep idlist$i.txt --freq --out $i | * Perform frequency analysis and association analysis using PLINK.<br>* 500 .freq and .assoc files will be created during this process |

| | | |
|---|---|---|
| | plink --bfile final.merged --keep idlist$i.txt --assoc --out $i<br>@ i++<br>end<br>exi | |
| 6 | **% freq.sh**<br>#!/bin/tcsh  –vx<br>awk '{print $1, $2, $3, $4}' 1.frq > name.txt<br>awk '{print $4}' mergethree.bim > name2.txt<br>paste name.txt name2.txt > MAF0.txt<br>awk '{print $1, $2, $3, $4}' 1.frq > N0.txt<br><br>@ i = 1<br>while ($i<= $1)<br>awk '{print $5}' $i.frq > MAF$i.txt<br>awk '{print $6}' $i.frq > N$i.txt<br>@ i++ | * Create the minor allele frequency and non-missing count files from all 500 replicates |

| | | |
|---|---|---|
| | End<br>paste MAF0.txt MAF1.txt MAF2.txt MAF3.txt MAF4.txt MAF5.txt MAF6.txt MAF7.txt MAF8.txt MAF9.txt MAF10.txt MAF11.txt MAF12.txt MAF13.txt MAF14.txt MAF15.txt MAF16.txt MAF17.txt MAF18.txt MAF19.txt MAF20.txt MAF21.txt MAF22.txt MAF23.txt MAF24.txt MAF25.txt MAF26.txt MAF27.txt MAF28.txt MAF29.txt MAF30.txt MAF31.txt MAF32.txt MAF33.txt MAF34.txt MAF35.txt MAF36.txt MAF37.txt MAF38.txt MAF39.txt MAF40.txt MAF41.txt MAF42.txt MAF43.txt MAF44.txt MAF45.txt MAF46.txt MAF47.txt MAF48.txt MAF49.txt MAF50.txt MAF51.txt MAF52.txt MAF53.txt MAF54.txt MAF55.txt MAF56.txt MAF57.txt MAF58.txt MAF59.txt MAF60.txt MAF61.txt MAF62.txt MAF63.txt MAF64.txt MAF65.txt MAF66.txt MAF67.txt MAF68.txt MAF69.txt MAF70.txt MAF71.txt MAF72.txt MAF73.txt MAF74.txt MAF75.txt MAF76.txt MAF77.txt MAF78.txt MAF79.txt MAF80.txt MAF81.txt MAF82.txt MAF83.txt MAF84.txt MAF85.txt MAF86.txt MAF87.txt MAF88.txt MAF89.txt MAF90.txt MAF91.txt MAF92.txt MAF93.txt MAF94.txt MAF95.txt MAF96.txt MAF97.txt | * Paste all minor allele frequency columns from each .freq files into the new files |

| | | |
|---|---|---|
| | MAF98.txt MAF99.txt MAF100.txt > tmp1.txt<br>paste MAF101.txt MAF102.txt MAF103.txt MAF104.txt MAF105.txt MAF106.txt MAF107.txt MAF108.txt MAF109.txt MAF110.txt MAF111.txt MAF112.txt MAF113.txt MAF114.txt MAF115.txt MAF116.txt MAF117.txt MAF118.txt MAF119.txt MAF120.txt MAF121.txt MAF122.txt MAF123.txt MAF124.txt MAF125.txt MAF126.txt MAF127.txt MAF128.txt MAF129.txt MAF130.txt MAF131.txt MAF132.txt MAF133.txt MAF134.txt MAF135.txt MAF136.txt MAF137.txt MAF138.txt MAF139.txt MAF140.txt MAF141.txt MAF142.txt MAF143.txt MAF144.txt MAF145.txt MAF146.txt MAF147.txt MAF148.txt MAF149.txt MAF150.txt MAF151.txt MAF152.txt MAF153.txt MAF154.txt MAF155.txt MAF156.txt MAF157.txt MAF158.txt MAF159.txt MAF160.txt MAF161.txt MAF162.txt MAF163.txt MAF164.txt MAF165.txt MAF166.txt MAF167.txt MAF168.txt MAF169.txt MAF170.txt MAF171.txt MAF172.txt MAF173.txt MAF174.txt MAF175.txt MAF176.txt MAF177.txt MAF178.txt MAF179.txt MAF180.txt MAF181.txt MAF182.txt MAF183.txt MAF184.txt MAF185.txt MAF186.txt MAF187.txt MAF188.txt MAF189.txt MAF190.txt | |

| | | |
|---|---|---|
| | MAF191.txt MAF192.txt MAF193.txt MAF194.txt MAF195.txt MAF196.txt MAF197.txt MAF198.txt MAF199.txt MAF200.txt > tmp2.txt<br>paste MAF201.txt MAF202.txt MAF203.txt MAF204.txt MAF205.txt MAF206.txt MAF207.txt MAF208.txt MAF209.txt MAF210.txt MAF211.txt MAF212.txt MAF213.txt MAF214.txt MAF215.txt MAF216.txt MAF217.txt MAF218.txt MAF219.txt MAF220.txt MAF221.txt MAF222.txt MAF223.txt MAF224.txt MAF225.txt MAF226.txt MAF227.txt MAF228.txt MAF229.txt MAF230.txt MAF231.txt MAF232.txt MAF233.txt MAF234.txt MAF235.txt MAF236.txt MAF237.txt MAF238.txt MAF239.txt MAF240.txt MAF241.txt MAF242.txt MAF243.txt MAF244.txt MAF245.txt MAF246.txt MAF247.txt MAF248.txt MAF249.txt MAF250.txt MAF251.txt MAF252.txt MAF253.txt MAF254.txt MAF255.txt MAF256.txt MAF257.txt MAF258.txt MAF259.txt MAF260.txt MAF261.txt MAF262.txt MAF263.txt MAF264.txt MAF265.txt MAF266.txt MAF267.txt MAF268.txt MAF269.txt MAF270.txt MAF271.txt MAF272.txt MAF273.txt MAF274.txt MAF275.txt MAF276.txt MAF277.txt MAF278.txt MAF279.txt MAF280.txt MAF281.txt MAF282.txt MAF283.txt MAF284.txt | |

| | | |
|---|---|---|
| | MAF285.txt MAF286.txt MAF287.txt MAF288.txt MAF289.txt MAF290.txt MAF291.txt MAF292.txt MAF293.txt MAF294.txt MAF295.txt MAF296.txt MAF297.txt MAF298.txt MAF299.txt MAF300.txt > tmp3.txt<br>paste MAF301.txt MAF302.txt MAF303.txt MAF304.txt MAF305.txt MAF306.txt MAF307.txt MAF308.txt MAF309.txt MAF310.txt MAF311.txt MAF312.txt MAF313.txt MAF314.txt MAF315.txt MAF316.txt MAF317.txt MAF318.txt MAF319.txt MAF320.txt MAF321.txt MAF322.txt MAF323.txt MAF324.txt MAF325.txt MAF326.txt MAF327.txt MAF328.txt MAF329.txt MAF330.txt MAF331.txt MAF332.txt MAF333.txt MAF334.txt MAF335.txt MAF336.txt MAF337.txt MAF338.txt MAF339.txt MAF340.txt MAF341.txt MAF342.txt MAF343.txt MAF344.txt MAF345.txt MAF346.txt MAF347.txt MAF348.txt MAF349.txt MAF350.txt MAF351.txt MAF352.txt MAF353.txt MAF354.txt MAF355.txt MAF356.txt MAF357.txt MAF358.txt MAF359.txt MAF360.txt MAF361.txt MAF362.txt MAF363.txt MAF364.txt MAF365.txt MAF366.txt MAF367.txt MAF368.txt MAF369.txt MAF370.txt MAF371.txt MAF372.txt MAF373.txt MAF374.txt MAF375.txt MAF376.txt MAF377.txt MAF378.txt | |

| | | |
|---|---|---|
| | MAF379.txt MAF380.txt MAF381.txt MAF382.txt MAF383.txt MAF384.txt MAF385.txt MAF386.txt MAF387.txt MAF388.txt MAF389.txt MAF390.txt MAF391.txt MAF392.txt MAF393.txt MAF394.txt MAF395.txt MAF396.txt MAF397.txt MAF398.txt MAF399.txt MAF400.txt >tmp4.txt<br>paste MAF401.txt MAF402.txt MAF403.txt MAF404.txt MAF405.txt MAF406.txt MAF407.txt MAF408.txt MAF409.txt MAF410.txt MAF411.txt MAF412.txt MAF413.txt MAF414.txt MAF415.txt MAF416.txt MAF417.txt MAF418.txt MAF419.txt MAF420.txt MAF421.txt MAF422.txt MAF423.txt MAF424.txt MAF425.txt MAF426.txt MAF427.txt MAF428.txt MAF429.txt MAF430.txt MAF431.txt MAF432.txt MAF433.txt MAF434.txt MAF435.txt MAF436.txt MAF437.txt MAF438.txt MAF439.txt MAF440.txt MAF441.txt MAF442.txt MAF443.txt MAF444.txt MAF445.txt MAF446.txt MAF447.txt MAF448.txt MAF449.txt MAF450.txt MAF451.txt MAF452.txt MAF453.txt MAF454.txt MAF455.txt MAF456.txt MAF457.txt MAF458.txt MAF459.txt MAF460.txt MAF461.txt MAF462.txt MAF463.txt MAF464.txt MAF465.txt MAF466.txt MAF467.txt MAF468.txt MAF469.txt MAF470.txt MAF471.txt MAF472.txt | |

| | | |
|---|---|---|
| | MAF473.txt MAF474.txt MAF475.txt MAF476.txt MAF477.txt MAF478.txt MAF479.txt MAF480.txt MAF481.txt MAF482.txt MAF483.txt MAF484.txt MAF485.txt MAF486.txt MAF487.txt MAF488.txt MAF489.txt MAF490.txt MAF491.txt MAF492.txt MAF493.txt MAF494.txt MAF495.txt MAF496.txt MAF497.txt MAF498.txt MAF499.txt MAF500.txt > tmp5.txt<br>paste tmp1.txt tmp2.txt tmp3.txt tmp4.txt tmp5.txt > MAF.txt<br>rm tmp1.txt<br>rm tmp2.txt<br>rm tmp3.txt<br>rm tmp4.txt<br>rm tmp5.txt | |
| | paste N0.txt N1.txt N2.txt N3.txt N4.txt N5.txt N6.txt N7.txt N8.txt N9.txt N10.txt N11.txt N12.txt N13.txt N14.txt N15.txt N16.txt N17.txt N18.txt N19.txt N20.txt N21.txt N22.txt N23.txt N24.txt N25.txt N26.txt N27.txt N28.txt N29.txt N30.txt N31.txt N32.txt N33.txt N34.txt N35.txt N36.txt N37.txt N38.txt N39.txt N40.txt N41.txt N42.txt N43.txt N44.txt N45.txt N46.txt N47.txt N48.txt N49.txt N50.txt | * Paste all non-missing allele count columns from each .freq files |

| | | |
|---|---|---|
| | N51.txt N52.txt N53.txt N54.txt N55.txt N56.txt N57.txt N58.txt N59.txt N60.txt N61.txt N62.txt N63.txt N64.txt N65.txt N66.txt N67.txt N68.txt N69.txt N70.txt N71.txt N72.txt N73.txt N74.txt N75.txt N76.txt N77.txt N78.txt N79.txt N80.txt N81.txt N82.txt N83.txt N84.txt N85.txt N86.txt N87.txt N88.txt N89.txt N90.txt N91.txt N92.txt N93.txt N94.txt N95.txt N96.txt N97.txt N98.txt N99.txt N100.txt > temp1.txt<br>paste N101.txt N102.txt N103.txt N104.txt N105.txt N106.txt N107.txt N108.txt N109.txt N110.txt N111.txt N112.txt N113.txt N114.txt N115.txt N116.txt N117.txt N118.txt N119.txt N120.txt N121.txt N122.txt N123.txt N124.txt N125.txt N126.txt N127.txt N128.txt N129.txt N130.txt N131.txt N132.txt N133.txt N134.txt N135.txt N136.txt N137.txt N138.txt N139.txt N140.txt N141.txt N142.txt N143.txt N144.txt N145.txt N146.txt N147.txt N148.txt N149.txt N150.txt N151.txt N152.txt N153.txt N154.txt N155.txt N156.txt N157.txt N158.txt N159.txt N160.txt N161.txt N162.txt N163.txt N164.txt N165.txt N166.txt N167.txt N168.txt N169.txt N170.txt N171.txt N172.txt N173.txt N174.txt N175.txt N176.txt N177.txt N178.txt N179.txt N180.txt N181.txt N182.txt N183.txt N184.txt N185.txt N186.txt N187.txt N188.txt N189.txt | |

| | | |
|---|---|---|
| | N190.txt N191.txt N192.txt N193.txt N194.txt N195.txt N196.txt N197.txt N198.txt N199.txt N200.txt > temp2.txt<br>paste N201.txt N202.txt N203.txt N204.txt N205.txt N206.txt N207.txt N208.txt N209.txt N210.txt N211.txt N212.txt N213.txt N214.txt N215.txt N216.txt N217.txt N218.txt N219.txt N220.txt N221.txt N222.txt N223.txt N224.txt N225.txt N226.txt N227.txt N228.txt N229.txt N230.txt N231.txt N232.txt N233.txt N234.txt N235.txt N236.txt N237.txt N238.txt N239.txt N240.txt N241.txt N242.txt N243.txt N244.txt N245.txt N246.txt N247.txt N248.txt N249.txt N250.txt N251.txt N252.txt N253.txt N254.txt N255.txt N256.txt N257.txt N258.txt N259.txt N260.txt N261.txt N262.txt N263.txt N264.txt N265.txt N266.txt N267.txt N268.txt N269.txt N270.txt N271.txt N272.txt N273.txt N274.txt N275.txt N276.txt N277.txt N278.txt N279.txt N280.txt N281.txt N282.txt N283.txt N284.txt N285.txt N286.txt N287.txt N288.txt N289.txt N290.txt N291.txt N292.txt N293.txt N294.txt N295.txt N296.txt N297.txt N298.txt N299.txt N300.txt > temp3.txt<br>paste N301.txt N302.txt N303.txt N304.txt N305.txt N306.txt N307.txt N308.txt N309.txt N310.txt N311.txt N312.txt N313.txt N314.txt N315.txt N316.txt N317.txt | |

| | | |
|---|---|---|
| | N318.txt N319.txt N320.txt N321.txt N322.txt N323.txt N324.txt N325.txt N326.txt N327.txt N328.txt N329.txt N330.txt N331.txt N332.txt N333.txt N334.txt N335.txt N336.txt N337.txt N338.txt N339.txt N340.txt N341.txt N342.txt N343.txt N344.txt N345.txt N346.txt N347.txt N348.txt N349.txt N350.txt N351.txt N352.txt N353.txt N354.txt N355.txt N356.txt N357.txt N358.txt N359.txt N360.txt N361.txt N362.txt N363.txt N364.txt N365.txt N366.txt N367.txt N368.txt N369.txt N370.txt N371.txt N372.txt N373.txt N374.txt N375.txt N376.txt N377.txt N378.txt N379.txt N380.txt N381.txt N382.txt N383.txt N384.txt N385.txt N386.txt N387.txt N388.txt N389.txt N390.txt N391.txt N392.txt N393.txt N394.txt N395.txt N396.txt N397.txt N398.txt N399.txt N400.txt > temp4.txt<br>paste N401.txt N402.txt N403.txt N404.txt N405.txt N406.txt N407.txt N408.txt N409.txt N410.txt N411.txt N412.txt N413.txt N414.txt N415.txt N416.txt N417.txt N418.txt N419.txt N420.txt N421.txt N422.txt N423.txt N424.txt N425.txt N426.txt N427.txt N428.txt N429.txt N430.txt N431.txt N432.txt N433.txt N434.txt N435.txt N436.txt N437.txt N438.txt N439.txt N440.txt N441.txt N442.txt N443.txt N444.txt N445.txt N446.txt N447.txt N448.txt N449.txt N450.txt N451.txt N452.txt N453.txt | |

| | | |
|---|---|---|
| | N454.txt N455.txt N456.txt N457.txt N458.txt N459.txt N460.txt N461.txt N462.txt N463.txt N464.txt N465.txt N466.txt N467.txt N468.txt N469.txt N470.txt N471.txt N472.txt N473.txt N474.txt N475.txt N476.txt N477.txt N478.txt N479.txt N480.txt N481.txt N482.txt N483.txt N484.txt N485.txt N486.txt N487.txt N488.txt N489.txt N490.txt N491.txt N492.txt N493.txt N494.txt N495.txt N496.txt N497.txt N498.txt N499.txt N500.txt > temp5.txt<br>paste temp1.txt temp2.txt temp3.txt temp4.txt temp5.txt > N.txt<br>rm temp1.txt<br>rm temp2.txt<br>rm temp3.txt<br>rm temp4.txt<br>rm temp5.txt<br>sed '/^0/d' MAF.txt \| sed '/^25/d' > mafall.txt<br>sed '/^0/d' N.txt \| sed '/^25/d' > nall.txt<br>@ i = 1<br>while ($i <= 22) | * Delete SNP sites that located out of chromosome 1~22 |

|  |  |  |
|---|---|---|
|  | grep "^$i " MAF.txt > mafch$i.txt<br>grep "^$i " N.txt > nch$i.txt<br>@ i+=1<br>end<br>exit | * Separate MAF and non-missing count files by different chromosomes |
| 7 | **% association.sh**<br>#!/bin/tcsh -vx<br>paste ASSOC0.txt ASSOC1.txt ASSOC2.txt ASSOC3.txt ASSOC4.txt ASSOC5.txt ASSOC6.txt ASSOC7.txt ASSOC8.txt ASSOC9.txt ASSOC10.txt ASSOC11.txt ASSOC12.txt ASSOC13.txt ASSOC14.txt ASSOC15.txt ASSOC16.txt ASSOC17.txt ASSOC18.txt ASSOC19.txt ASSOC20.txt ASSOC21.txt ASSOC22.txt ASSOC23.txt ASSOC24.txt ASSOC25.txt ASSOC26.txt ASSOC27.txt ASSOC28.txt ASSOC29.txt ASSOC30.txt ASSOC31.txt ASSOC32.txt ASSOC33.txt ASSOC34.txt ASSOC35.txt ASSOC36.txt ASSOC37.txt ASSOC38.txt ASSOC39.txt ASSOC40.txt ASSOC41.txt ASSOC42.txt ASSOC43.txt ASSOC44.txt ASSOC45.txt ASSOC46.txt ASSOC47.txt ASSOC48.txt ASSOC49.txt ASSOC50.txt ASSOC51.txt ASSOC52.txt ASSOC53.txt | * Create association p-value lists from all 500 .assoc siles |

| | | |
|---|---|---|
| | ASSOC54.txt ASSOC55.txt ASSOC56.txt ASSOC57.txt ASSOC58.txt ASSOC59.txt ASSOC60.txt ASSOC61.txt ASSOC62.txt ASSOC63.txt ASSOC64.txt ASSOC65.txt ASSOC66.txt ASSOC67.txt ASSOC68.txt ASSOC69.txt ASSOC70.txt ASSOC71.txt ASSOC72.txt ASSOC73.txt ASSOC74.txt ASSOC75.txt ASSOC76.txt ASSOC77.txt ASSOC78.txt ASSOC79.txt ASSOC80.txt ASSOC81.txt ASSOC82.txt ASSOC83.txt ASSOC84.txt ASSOC85.txt ASSOC86.txt ASSOC87.txt ASSOC88.txt ASSOC89.txt ASSOC90.txt ASSOC91.txt ASSOC92.txt ASSOC93.txt ASSOC94.txt ASSOC95.txt ASSOC96.txt ASSOC97.txt ASSOC98.txt ASSOC99.txt ASSOC100.txt > tmp1.txt<br>paste ASSOC101.txt ASSOC102.txt ASSOC103.txt ASSOC104.txt ASSOC105.txt ASSOC106.txt ASSOC107.txt ASSOC108.txt ASSOC109.txt ASSOC110.txt ASSOC111.txt ASSOC112.txt ASSOC113.txt ASSOC114.txt ASSOC115.txt ASSOC116.txt ASSOC117.txt ASSOC118.txt ASSOC119.txt ASSOC120.txt ASSOC121.txt ASSOC122.txt ASSOC123.txt ASSOC124.txt ASSOC125.txt ASSOC126.txt ASSOC127.txt ASSOC128.txt ASSOC129.txt ASSOC130.txt ASSOC131.txt ASSOC132.txt ASSOC133.txt ASSOC134.txt ASSOC135.txt ASSOC136.txt ASSOC137.txt ASSOC138.txt ASSOC139.txt ASSOC140.txt | |

| | | |
|---|---|---|
| | ASSOC141.txt ASSOC142.txt ASSOC143.txt ASSOC144.txt ASSOC145.txt ASSOC146.txt ASSOC147.txt ASSOC148.txt ASSOC149.txt ASSOC150.txt ASSOC151.txt ASSOC152.txt ASSOC153.txt ASSOC154.txt ASSOC155.txt ASSOC156.txt ASSOC157.txt ASSOC158.txt ASSOC159.txt ASSOC160.txt ASSOC161.txt ASSOC162.txt ASSOC163.txt ASSOC164.txt ASSOC165.txt ASSOC166.txt ASSOC167.txt ASSOC168.txt ASSOC169.txt ASSOC170.txt ASSOC171.txt ASSOC172.txt ASSOC173.txt ASSOC174.txt ASSOC175.txt ASSOC176.txt ASSOC177.txt ASSOC178.txt ASSOC179.txt ASSOC180.txt ASSOC181.txt ASSOC182.txt ASSOC183.txt ASSOC184.txt ASSOC185.txt ASSOC186.txt ASSOC187.txt ASSOC188.txt ASSOC189.txt ASSOC190.txt ASSOC191.txt ASSOC192.txt ASSOC193.txt ASSOC194.txt ASSOC195.txt ASSOC196.txt ASSOC197.txt ASSOC198.txt ASSOC199.txt ASSOC200.txt > tmp2.txt<br>paste ASSOC201.txt ASSOC202.txt ASSOC203.txt ASSOC204.txt ASSOC205.txt ASSOC206.txt ASSOC207.txt ASSOC208.txt ASSOC209.txt ASSOC210.txt ASSOC211.txt ASSOC212.txt ASSOC213.txt ASSOC214.txt ASSOC215.txt | |

| | | |
|---|---|---|
| | ASSOC216.txt ASSOC217.txt ASSOC218.txt ASSOC219.txt ASSOC220.txt ASSOC221.txt ASSOC222.txt ASSOC223.txt ASSOC224.txt ASSOC225.txt ASSOC226.txt ASSOC227.txt ASSOC228.txt ASSOC229.txt ASSOC230.txt ASSOC231.txt ASSOC232.txt ASSOC233.txt ASSOC234.txt ASSOC235.txt ASSOC236.txt ASSOC237.txt ASSOC238.txt ASSOC239.txt ASSOC240.txt ASSOC241.txt ASSOC242.txt ASSOC243.txt ASSOC244.txt ASSOC245.txt ASSOC246.txt ASSOC247.txt ASSOC248.txt ASSOC249.txt ASSOC250.txt ASSOC251.txt ASSOC252.txt ASSOC253.txt ASSOC254.txt ASSOC255.txt ASSOC256.txt ASSOC257.txt ASSOC258.txt ASSOC259.txt ASSOC260.txt ASSOC261.txt ASSOC262.txt ASSOC263.txt ASSOC264.txt ASSOC265.txt ASSOC266.txt ASSOC267.txt ASSOC268.txt ASSOC269.txt ASSOC270.txt ASSOC271.txt ASSOC272.txt ASSOC273.txt ASSOC274.txt ASSOC275.txt ASSOC276.txt ASSOC277.txt ASSOC278.txt ASSOC279.txt ASSOC280.txt ASSOC281.txt ASSOC282.txt ASSOC283.txt ASSOC284.txt ASSOC285.txt ASSOC286.txt ASSOC287.txt ASSOC288.txt ASSOC289.txt ASSOC290.txt ASSOC291.txt ASSOC292.txt ASSOC293.txt ASSOC294.txt ASSOC295.txt | |

| | | |
|---|---|---|
| | ASSOC296.txt ASSOC297.txt ASSOC298.txt ASSOC299.txt ASSOC300.txt > tmp3.txt<br>paste ASSOC301.txt ASSOC302.txt ASSOC303.txt ASSOC304.txt ASSOC305.txt ASSOC306.txt ASSOC307.txt ASSOC308.txt ASSOC309.txt ASSOC310.txt ASSOC311.txt ASSOC312.txt ASSOC313.txt ASSOC314.txt ASSOC315.txt ASSOC316.txt ASSOC317.txt ASSOC318.txt ASSOC319.txt ASSOC320.txt ASSOC321.txt ASSOC322.txt ASSOC323.txt ASSOC324.txt ASSOC325.txt ASSOC326.txt ASSOC327.txt ASSOC328.txt ASSOC329.txt ASSOC330.txt ASSOC331.txt ASSOC332.txt ASSOC333.txt ASSOC334.txt ASSOC335.txt ASSOC336.txt ASSOC337.txt ASSOC338.txt ASSOC339.txt ASSOC340.txt ASSOC341.txt ASSOC342.txt ASSOC343.txt ASSOC344.txt ASSOC345.txt ASSOC346.txt ASSOC347.txt ASSOC348.txt ASSOC349.txt ASSOC350.txt ASSOC351.txt ASSOC352.txt ASSOC353.txt ASSOC354.txt ASSOC355.txt ASSOC356.txt ASSOC357.txt ASSOC358.txt ASSOC359.txt ASSOC360.txt ASSOC361.txt ASSOC362.txt ASSOC363.txt ASSOC364.txt ASSOC365.txt ASSOC366.txt ASSOC367.txt ASSOC368.txt ASSOC369.txt ASSOC370.txt | |

| | | |
|---|---|---|
| | ASSOC371.txt ASSOC372.txt ASSOC373.txt ASSOC374.txt ASSOC375.txt ASSOC376.txt ASSOC377.txt ASSOC378.txt ASSOC379.txt ASSOC380.txt ASSOC381.txt ASSOC382.txt ASSOC383.txt ASSOC384.txt ASSOC385.txt ASSOC386.txt ASSOC387.txt ASSOC388.txt ASSOC389.txt ASSOC390.txt ASSOC391.txt ASSOC392.txt ASSOC393.txt ASSOC394.txt ASSOC395.txt ASSOC396.txt ASSOC397.txt ASSOC398.txt ASSOC399.txt ASSOC400.txt >tmp4.txt<br>paste ASSOC401.txt ASSOC402.txt ASSOC403.txt ASSOC404.txt ASSOC405.txt ASSOC406.txt ASSOC407.txt ASSOC408.txt ASSOC409.txt ASSOC410.txt ASSOC411.txt ASSOC412.txt ASSOC413.txt ASSOC414.txt ASSOC415.txt ASSOC416.txt ASSOC417.txt ASSOC418.txt ASSOC419.txt ASSOC420.txt ASSOC421.txt ASSOC422.txt ASSOC423.txt ASSOC424.txt ASSOC425.txt ASSOC426.txt ASSOC427.txt ASSOC428.txt ASSOC429.txt ASSOC430.txt ASSOC431.txt ASSOC432.txt ASSOC433.txt ASSOC434.txt ASSOC435.txt ASSOC436.txt ASSOC437.txt ASSOC438.txt ASSOC439.txt ASSOC440.txt ASSOC441.txt ASSOC442.txt ASSOC443.txt ASSOC444.txt ASSOC445.txt | |

| | | |
|---|---|---|
| | ASSOC446.txt ASSOC447.txt ASSOC448.txt ASSOC449.txt ASSOC450.txt<br>ASSOC451.txt ASSOC452.txt ASSOC453.txt ASSOC454.txt ASSOC455.txt<br>ASSOC456.txt ASSOC457.txt ASSOC458.txt ASSOC459.txt ASSOC460.txt<br>ASSOC461.txt ASSOC462.txt ASSOC463.txt ASSOC464.txt ASSOC465.txt<br>ASSOC466.txt ASSOC467.txt ASSOC468.txt ASSOC469.txt ASSOC470.txt<br>ASSOC471.txt ASSOC472.txt ASSOC473.txt ASSOC474.txt ASSOC475.txt<br>ASSOC476.txt ASSOC477.txt ASSOC478.txt ASSOC479.txt ASSOC480.txt<br>ASSOC481.txt ASSOC482.txt ASSOC483.txt ASSOC484.txt ASSOC485.txt<br>ASSOC486.txt ASSOC487.txt ASSOC488.txt ASSOC489.txt ASSOC490.txt<br>ASSOC491.txt ASSOC492.txt ASSOC493.txt ASSOC494.txt ASSOC495.txt<br>ASSOC496.txt ASSOC497.txt ASSOC498.txt ASSOC499.txt ASSOC500.txt ><br>tmp5.txt<br>paste tmp1.txt tmp2.txt tmp3.txt tmp4.txt tmp5.txt > ASSOC.txt<br>rm tmp1.txt<br>rm tmp2.txt<br>rm tmp3.txt | |

| | | |
|---|---|---|
| | rm tmp4.txt<br>rm tmp5.txt<br>end<br>exit | |
| 8 | **% location.sh**<br>#!/bin/tcsh –vx<br>R CMD BATCH SD.R<br>@ i = 1<br>while ($i<= 22)<br>awk '{print $5}' mafch$i.txt > loca$i.txt<br>paste sd$i.txt loca$i.txt > sd1$i.txt<br>@ i++<br>R CMD BATCH BIN.R<br>R CMD BATCH plot.R<br>R CMD BATCH  associate.R<br>end | * Create data lists with standard deviation of 500 MAF replicates and location (bp) for each SNP site.<br>* Create data lists with bin number |

|  |  |  |
|---|---|---|
|  | exit |  |
| 9 | **% BIN.R**<br>for (j in 1:22)<br>{<br>c <- rep(0,nrow(sd1[[j]]))<br>for (i in 1:length(c))<br>{<br>if (sd1[[j]]$V4[i]>1 & sd1[[j]]$V4[i]<10000000)  { c[i] <- 1 } else<br>if (sd1[[j]]$V4[i]>10000001 & sd1[[j]]$V4[i]<20000000) { c[i] <-2} else<br>if (sd1[[j]]$V4[i]>20000001 & sd1[[j]]$V4[i]<30000000) {c[i] <-3} else<br>if (sd1[[j]]$V4[i]>30000001 & sd1[[j]]$V4[i]<40000000) { c[i] <-4} else<br>if (sd1[[j]]$V4[i]>40000001 & sd1[[j]]$V4[i]<50000000) { c[i] <-5} else<br>if (sd1[[j]]$V4[i]>50000001 & sd1[[j]]$V4[i]<60000000) { c[i] <-6} else<br>if (sd1[[j]]$V4[i]>60000001 & sd1[[j]]$V4[i]<70000000) { c[i] <-7} else<br>if (sd1[[j]]$V4[i]>70000001 & sd1[[j]]$V4[i]<80000000) { c[i] <-8} else<br>if (sd1[[j]]$V4[i]>80000001 & sd1[[j]]$V4[i]<90000000) { c[i] <-9} else | * Create bin numbers using the SNP location. SNPs will be grouped every 10,000,000 bp. The longest chromosome, chromosome 1 and 2, consists 25 bins each, where the shortest one, chromosome 21 and 22, consists 5 bins each.<br>* sd1[i]$V4 contains the location (bp) of each SNP site |

```
if (sd1[[j]]$V4[i]>90000001 & sd1[[j]]$V4[i]<100000000) { c[i] <-10} else

if (sd1[[j]]$V4[i]>100000001 & sd1[[j]]$V4[i]<110000000) { c[i] <-11} else

if (sd1[[j]]$V4[i]>110000001 & sd1[[j]]$V4[i]<120000000) { c[i] <-12} else

if (sd1[[j]]$V4[i]>120000001 & sd1[[j]]$V4[i]<130000000) { c[i] <-13} else

if (sd1[[j]]$V4[i]>130000001 & sd1[[j]]$V4[i]<140000000) { c[i] <-14} else

if (sd1[[j]]$V4[i]>140000001 & sd1[[j]]$V4[i]<150000000) { c[i] <-15} else

if (sd1[[j]]$V4[i]>150000001 & sd1[[j]]$V4[i]<160000000) { c[i] <-16} else

if (sd1[[j]]$V4[i]>160000001 & sd1[[j]]$V4[i]<170000000) { c[i] <-17} else

if (sd1[[j]]$V4[i]>170000001 & sd1[[j]]$V4[i]<180000000) { c[i] <-18} else

if (sd1[[j]]$V4[i]>180000001 & sd1[[j]]$V4[i]<190000000) { c[i] <-19} else

if (sd1[[j]]$V4[i]>190000001 & sd1[[j]]$V4[i]<200000000) { c[i] <-20} else

if (sd1[[j]]$V4[i]>200000001 & sd1[[j]]$V4[i]<210000000) { c[i] <-21} else

if (sd1[[j]]$V4[i]>210000001 & sd1[[j]]$V4[i]<220000000) { c[i] <-22} else

if (sd1[[j]]$V4[i]>220000001 & sd1[[j]]$V4[i]<230000000) { c[i] <-23} else

if (sd1[[j]]$V4[i]>230000001 & sd1[[j]]$V4[i]<240000000) { c[i] <-24} else

if (sd1[[j]]$V4[i]>240000001 & sd1[[j]]$V4[i]<250000000) { c[i] <-25}
```

|  |  |  |
|---|---|---|
|  | }<br>write.table(c,paste("bin",i,".txt",sep="")sep="\t",row.names = FALSE,col.names=FALSE, quote = FALSE)<br>} |  |
| 10 | **% SD.R**<br>tmp<-read.csv("num.tmp",header=F)<br>selection <- 2*tmp[1,2]*0.5<br>fname <- paste("mafch",1:22,".txt",sep="")<br>nname <- paste("nch",1:22,".txt",sep="")<br>sitename <- paste("site",1:22,".txt",sep="")<br>sd <- vector("list",length=22)<br>for (i in 1:22)<br>{<br>maf<- read.table(fname[i],sep="",head=F)<br>n<- read.table(nname[i],sep="",head=F)<br>site <- read.table(sitename[i],sep="",head=F) | * Read minor allele frequency and non-missing count data by chromosome, using proper selection way to calculate the standard deviation of 500 MAF replicates for each SNP site.<br>* The result will end up with sd1.txt to sd22.txt files. |

|  |  |  |
|---|---|---|
|  | n2 <- t(apply(n[,5:504],1,function(x)ifelse(x<selection,NA,x)))<br>maf[,6:505][is.na(n2)] <-NA<br>sd[[i]] <- apply(maf[,6:505],1,sd,na.rm=T)<br>sdd<- cbind(site,sd[[i]])<br>rm(maf)<br>rm(n)<br>rm(n2)<br>rm(site)<br>write.table(sdd, paste("sd",i,".txt",sep=""),sep="\t",row.names = FALSE,col.names= FALSE, quote = FALSE)<br>} |  |
| 11 | **% plot.R**<br>sd1 <- vector("list",length=22)<br>for (i in 1:22)<br>{sd1[[i]] <- read.table(paste("sd",i,".txt",sep=""),sep="\t",head=F)}<br>eachchromean <- vector("list",22) | * Read in all the sd files by chromosomes<br>* sd1[i]$V3 contains the standard deviation of 500 minor allele frequency replicates |

| | | |
|---|---|---|
| | for (i in 1:22)<br>{eachchromean[[i]] <- mean(sd1[[i]]$V3, na.rm=T)}<br>meanplussd <- vector("list",22)<br>for (i in 1:22)<br>{meanplussd[[i]] <- mean(sd1[[i]]$V3, na.rm=T)+2*sd(sd1[[i]]$V3, na.rm=T)}<br>number <- c(1:22)<br>chromosd <- unlist(eachchromean)<br>cutpoint <- unlist(meanplussd)<br>chrmeansd <- t(cbind(number, chromosd))<br>write.table(chrmeansd, "chromosomesd.txt", sep="\t",row.names = FALSE, col.names= FALSE, quote = FALSE)<br>eachsd <- vector("list",22)<br>for (i in 1:22)<br>{<br>eachsd[[i]] <- sd1[[i]]$V3 | * sd1[i]$V4 contains the location (bp) of each SNP site |

```
}
kk <- c(1:22)
dim <- length(kk)
ttestmatrix <- matrix(" ", dim, dim)
var.pvalue <- matrix(" ", dim, dim)
for (i in 1: length(eachsd)){m <-eachsd[kk]}
for (j in 1: length( m)){
for ( l in 1: length( m)){
var.pvalue[j,l] <- var.test(m[[j]],m[[ l]])$p.value
if (var.pvalue[j,l] < 0.05) {
if ( kk[j] > kk[ l]) {
if (length(m[[j]])==1) ttestmatrix[j,] <- 0
else if (length(m[[ l]])==1) ttestmatrix[, l] <- 0
else ttestmatrix[j,l] <- t.test(m[[j]],m[[ l]], var.equal =F)$p.value
}}
else if (var.pvalue[j,l] >= 0.05)
```

* t-test and equal variance test was performed to compare the average SD in each chromosome

| | | |
|---|---|---|
| | {if ( kk[j] > kk[ l]) {if (length(m[[j]])==1) ttestmatrix[j,] <- 0<br>else if (length(m[[ l]])==1) ttestmatrix[, l] <- 0<br>else ttestmatrix[j,l] <- t.test(m[[j]],m[[ l]], var.equal =T)$p.value<br>}}<br>}}<br>write.table(ttestmatrix, "ttest.csv",row.names = T, col.names= T, quote = FALSE) | |
| | bmp("chromowide.bmp", width=1600, height=800, pointsize=23)<br>par(mfrow=c(2,12))<br>for (i in 1: length(kk))<br>{boxplot(sd1[[i]]$V3, ylim=c(0,0.12),xlab=paste("chr",kk[i]))}<br>dev.off() | * Plot the boxplot of SD in each chromosome |
| | outlier <- vector("list",22)<br>for (i in 1:22)<br>{outlier[[i]] <- sd1[[i]][which(sd1[[i]]$V3>meanplussd[i]),]} | * Plot the dotplot of outliers in each chromosome |

|  |  |  |
|---|---|---|
|  | for (i in 1:22)<br>{<br>bmp(paste("outlier",i,".bmp"), height=600, width=1000, pointsize=12)<br>plot(x=outlier[[i]]$V4,y=outlier[[i]]$V3, type="p", xlim=c(0,2.5e+08), axes=FALSE, ylim=c(0.035,0.12), xlab= paste(""), ylab=paste(""))<br>axis(1,at=c(0, 1.0e+07, 2.0e+07, 3.0e+07, 4.0e+07, 5.0e+07, 6.0e+07, 7.0e+07, 8.0e+07, 9.0e+07, 1.0e+08, 1.1e+08, 1.2e+08, 1.3e+08, 1.4e+08, 1.5e+08, 1.6e+08, 1.7e+08, 1.8e+08, 1.9e+08, 2.0e+08, 2.1e+08, 2.2e+08, 2.3e+08, 2.4e+08, 2.5e+08), labels=c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25))<br>axis(2, at=c(0.03, 0.06, 0.09, 0.12), labels=c(0.03, 0.06, 0.09, 0.12))<br> dev.off()<br>} |  |
| 12 | **% associate.R**<br>pcriteria <- read.table("pcriteria.txt",header=F)<br>counting<-function(x) | * Using sample dependent p-value as criteria, ‘pcriteria.txt’ contains different p criteria for different |

| | | |
|---|---|---|
| | {<br>y<-x[!is.na(x)]<br>z<-length(y[y<pcritera[1,2]])<br>z<br>}<br>for (i in 1:22) | sample size, and will be changed from 0.05 to 0.0001 |
| | {<br>bin <- read.table(paste("bin",i,".txt",sep=""),sep="\t",na.strings="NA",head=T)<br>pvalues <- read.table(paste("assocall",i,".txt",sep=""),na.strings="NA",head=F)<br>pvalues1 <- as.matrix(pvalues[,3:502])<br>repnum <- apply(pvalues1,1,counting)<br>out1 <- cbind(rep(i,length(bin)),bin,repnum)<br>write.table(out1,paste("p",i,".txt",sep=""),sep="\t",row.names = FALSE,col.names= FALSE, quote = FALSE)<br>rm(bin)<br>rm(pvalues)<br>rm(pvalues1) | * Read in bin and p-value files for each chromosome<br>* Use the function above to count number of replicates that with significant p-value for each SNP site<br>* pi.txt file contains the number of significant p-value for each SNP site |

| | | |
|---|---|---|
| | rm(out1)<br>}<br>for (j in 1:22)<br>{<br>p <- read.table(paste("p",j,".txt",sep=""),sep="\t",head=F)<br>out <- NULL<br>for(i in 1:25)<br>{<br>out<-c(out,mean(p[p$V2==i,3],na.rm=T))<br>}<br>write.table(out,paste("pcount",j,".txt",sep=""),sep="\t",row.names = FALSE,col.names=<br>FALSE, quote = FALSE)<br>rm(p)<br>rm(out)<br>} | * Read in the pi.txt files<br>* 25 is the largest bin size (in chromosome 1 and 2), p$V2 is the column of bin number.<br>* I took the means of ‘number of sig. p-value’ for each bin, and write out the list ‘pcounti.txt’ for each chromosome |

# BIBLIOGRAPHY

1. Liew SH, Elsner H, Spector TD, Hammond CJ (2005) The first "classical" twin study? Analysis of refractive error using monozygotic and dizygotic twins published in 1922. Twin Res Hum Genet 8:198-200

2. Jablonski W (1922) Ein Beitrag zur Vererbung der Refraktion menschlicher Augen. [A contribution to the heredity of refraction in human eyes]. Arch Augenheilk 91:308-328

3. Siemens HW (1924) Zwillingspathologie: Ihre Bedeutung; ihre Methodik, ihre bisherigen Ergebnisse [Twin pathology: Its meaning; its method; results so far]. Berlin, Germany: Springer Verlag

4. Merrriman C (1924) The intellectual resemblance of twins. Psychological Monographs 33:1-58

5. Avery OT MC, and McCarty M (1944) Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. Journal of Experimental Medicine 79:137-158

6. Hershey AD, Chase M (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen Physiol 36:39-56

7. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171:737-738

8. Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503-517

9. Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci U S A 74:560-564

10. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463-5467

11. Ellegren H, Parsch J (2007) The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet 8:689-698

12. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453:56-64

13. Jorgenson E, Witte JS (2006) Coverage and power in genomewide association studies. Am J Hum Genet 78:884-888

14. Teo YY (2008) Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. Curr Opin Lipidol 19:133-143

15. Spencer CC, Su Z, Donnelly P, Marchini J (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet 5:e1000477

16. Klein RJ (2007) Power analysis for genome-wide association studies. BMC Genet 8:58

17. de Bakker PI, Burtt NP, Graham RR, Guiducci C, Yelensky R, Drake JA, Bersaglieri T, Penney KL, Butler J, Young S, Onofrio RC, Lyon HN, Stram DO, Haiman CA, Freedman ML, Zhu X, Cooper R, Groop L, Kolonel LN, Henderson BE, Daly MJ, Hirschhorn JN, Altshuler D (2006) Transferability of tag SNPs in genetic association studies in multiple populations. Nat Genet 38:1298-1303

18. Hao K, Chudin E, McElwee J, Schadt EE (2009) Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. BMC Genet 10:27

19. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, et al. (2007) The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 39:1181-1186

20. Boomsma DI, Willemsen G, Sullivan PF, Heutink P, Meijer P, Sondervan D, Kluft C, Smit G, Nolen WA, Zitman FG, Smit JH, Hoogendijk WJ, van Dyck R, de Geus EJ, Penninx BW (2008) Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. Eur J Hum Genet 16:335-342

21. Boomsma DI, Beem AL, van den Berg M, Dolan CV, Koopmans JR, Vink JM, de Geus EJ, Slagboom PE (2000) Netherlands twin family study of anxious depression (NETSAD). Twin Res 3:323-334

22. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G, Vollenweider P, Oksenberg JR, Hauser SL, Stirnadel HA, Kooner JS, Chambers JC, Jones B, Mooser V, Bustamante CD, Roses AD, Burns DK, Ehm MG, Lai EH (2008) The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83:347-358

23. Fung HC, Scholz S, Matarin M, Simon-Sanchez J, Hernandez D, Britton A, Gibbs JR, Langefeld C, Stiegert ML, Schymick J, Okun MS, Mandel RJ, Fernandez HH, Foote KD, Rodriguez RL, Peckham E, De Vrieze FW, Gwinn-Hardy K, Hardy JA, Singleton A (2006) Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol 5:911-916

24. (2003) The International HapMap Project. Nature 426:789-796

25. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381-2385

26. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, et al. (2002) A human genome diversity cell line panel. Science 296:261-262

27. Cavalli-Sforza LL (2005) The Human Genome Diversity Project: past, present and future. Nat Rev Genet 6:333-340

28. Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet 70:841-847

29. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575

30. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, Mefford H, Ying P, Nickerson DA, Eichler EE (2009) Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 84:148-161