Localized Correlation Analysis and Genetic Association with Cardiovascular Disease

by

Chern-Han Ou

Computational Biology & Bioinformatics Program
Duke University

Date:_____
Approved:

_____
Elizabeth Hauser, Supervisor

_____
Simon Gregory

_____
Jeanette McCarthy

_____
Sayan Mukherjee

Thesis submitted in partial fulfillment of
the requirements for the degree of Master of Science in the
Computational Biology & Bioinformatics Program in the Graduate School
of Duke University

2010

ABSTRACT

Localized Correlation Analysis and Genetic Association with Cardiovascular Disease

by

Chern-Han Ou

Computational Biology & Bioinformatics Program
Duke University

Date:_____
Approved:

_____
Elizabeth Hauser, Supervisor

_____
Simon Gregory

_____
Jeanette McCarthy

_____
Sayan Mukherjee

An abstract of a thesis submitted in partial
fulfillment of the requirements for the degree
of Master of Science in the
Computational Biology & Bioinformatics Program in the Graduate School
of Duke University

2010

Copyright by
Chern-Han Ou
2010

Abstract

Variations in gene expression are potential risk factors for atherosclerosis, which is one of the most common forms of cardiovascular disease. We performed a localized Pearson correlation test in 372 individuals from seven datasets relevant to cardiovascular disease studies. The genomes of samples were separated into 20Mb windows and correlation tests were performed locally in these windows. The localized Pearson correlation test found chr3:115Mb-135Mb was tightly connected by significantly high proportion of highly correlated pairs (*P* value = 0.0266 with Z-test). LSAMP, GATA2, MBD4, and other genes in the region were considered associated with cardiovascular disease because they were involved in highly correlated pairs. Furthermore, these genes were also associated with cardiovascular disease by having significantly high SNP odds ratios (*P* value < 0.1) between patients and controls in an independent Duke University Medical Center database. In addition, a permutation test was performed to demonstrate that chr3:115Mb-135Mb might underlie the regulation of cardiovascular disease. Finally, the localized Pearson correlation test also found some other regions that could be associated with cardiovascular disease.

Contents

Abstract iv

List of Tables vii

List of Figures viii

Acknowledgements xi

1. Introduction 1

1.1 Cardiovascular Disease 1

1.2 Gene Expression and Correlation Coefficients 5

1.3 Specific Aims 7

2. Materials and Methods 8

2.1 Subjects 8

2.2 Statistical Analysis 11

2.2.1 The Localized Pearson Correlation Test 11

2.2.2 Permutation Test 14

3. Results 16

3.1 Pair-wise correlation between 115Mb and 135Mb on chromosome 3 16

3.2 Characteristics of highly correlated genes from the Aorta dataset 18

3.3 GATA2 and LSAMP are at the center of the connections 19

3.4 115Mb – 135Mb on chromosome 3 is a region with significantly strong connectivity in whole genome 28

3.5 Some other regions in the genome showed significantly strong connectivity 33

3.6 Varying the threshold for the correlation coefficient 43

3.7 Analyzing 115Mb – 135Mb on chromosome 3 from CATHGEN dataset.....51

3.8 Analyzing 115Mb – 135Mb on chromosome 3 from GDS1597 dataset.....51

3.9 115Mb-135Mb on chromosome 3 showed more connections to the rest of the genome than others.....54

4. Discussion 58

4.1 115Mb-135Mb on Chromosome 3 58

4.2 The Regions Detected by the Localized Pearson's Correlation Test 67

4.3 Factor of different results might be Tissue-specific.....71

4.4 Conclusion.....72

Appendix A.....74

Appendix B 81

Appendix C.....82

References 83

List of Tables

Table 1: Clinical characteristics of the seven expression data sets. 9

Table 2: Proportion of the pairs with high correlation in the 115Mb-135Mb region on chromosome 3 compared across seven expression data sets. 17

Table 3: All regions with a significantly large proportion of highly correlated pairs (*R*) in the Aorta dataset. The sampling windows which overlapped with each other over 10 Mb were merged into a single region. The *P* value came from Z-test of proportion of highly correlated pairs. 34

Table 4: The list of genes appearing in highly correlated pairs. This table compares two lists of genes extracted from the Aorta dataset and the GDS1597 dataset. There are 20 genes appearing in both lists, 25 genes appear only in the gene list from Aorta dataset, and 51 genes appear only in the gene list from GDS1597 dataset. 53

Table 5: SNP-verification of genes involved in highly correlated pairs in the Aorta dataset or GDS1597 dataset. A gene was verified if there is at least one SNP with a significantly different (*P* value < 0.1) in proportion of minor allele in either the Aorta database or CATHGEN database. 60

Table 6: Genes with significantly different odd ratios between cardiovascular disease patients and controls in DUMC SNPs database. These genes were also participated in highly correlated pairs from the Aorta dataset. Most genes involved in highly correlated pairs could not be verified by SNPs since many genes have not been genotyped in the case-control series. 69

List of Figures

Figure 1: Definition of regions relative to expression tags. The red segments in figures show tags on a microarray. One strategy is that the size of the sampling region and the step size are based on a fixed length in basepairs. For example, in (a), the length of the sampling region is fixed at 20Mb. The other strategy is that the size of the sampling region and step size are based on a fixed number of tags. In (b), the length of the sampling region is fixed at 4 tags, for example. 13

Figure 2: Distribution of distances of physical position between highly correlated pairs in 115Mb to 135Mb on chromosome 3 from the Aorta dataset. The bar plot shows the frequency of distances between highly correlated pairs. The lower panel, the box plot, shows the range, quartile, and median of distances of the highly correlated pairs. 18

Figure 3: The frequencies of genes appearing in the high correlation pairs in 115Mb to 135Mb on chromosome 3 from the Aorta dataset. The genes are arranged by the order of physical locations. 20

Figure 4: Scatter plots of normalized gene expression values of GATA2 and LSAMP in the Aorta dataset. The correlation coefficient between GATA2 and LSAMP expressions is 0.7846. 21

Figure 5: Scatter plots of GATA2 and LSAMP with the genes that are highly correlated to both GATA2 and LSAMP in the Aorta data set. The correlations demonstrate a *cis*-directed property, except for HEG1. 22

Figure 6: Scatter plot of GATA2 and LSAMP in the CATHGEN dataset. In U133plus2 microarray platform, Affymetrix IDs of GATA2 are 207954_at, 209710_at, and 210358_x_at, and Affymetrix IDs of LSAMP are 1570190_at and 214460_at. The scatter plot suggested that there is no relationship between GATA2 and LSAMP in the CATHGEN dataset. 28

Figure 7: The cumulative distribution function of proportion of high correlation coverage, *R*, from the Aorta data set. The red dot represents value of chr3:115Mb-135Mb. The sampling windows included all possible *R* values without applying a filter for the number of tags. 29

Figure 8: Distribution of proportion of high correlation coverage, *R*, from Aorta dataset. (a) The regions with less than 18 tags are included in the distribution and

represent in red blocks. (b) The regions with less than 18 tags are filtered out. Many extreme values are eliminated and the distribution of *R* becomes closer to a normal distribution. (c) The distribution of *R* after applying tag-oriented sampling method. By fixing the number of tags in a window, the extreme values are eliminated. 31

Figure 9: The results of high correlation analysis for the genome applied to the Aorta dataset. The blue curves show the distributions of logarithm of p-value of every region, and the red curves represent the number of tags in the region. The physical position of a region represents the midpoint of the range. For example, the analysis result from sampling window 115Mb-135Mb was plotted on physical position 125Mb. 37

Figure 10: The statistical characteristics of the proportion of highly correlated pairs (*R*) resulted from various thresholds for the correlation coefficients from (a) the Aorta dataset and (b) the CATHGEN dataset. The *R* values from the regions with number of tags lower than 18 were eliminated. The red squares indicate the *R* values of 115Mb-135Mb on chromosome 3. 44

Figure 11: The distributions of proportion of highly correlated pairs of each chromosome from the Aorta dataset while varying threshold of correlation coefficient. The dotted curves represent the number of tags in the region for reference. 45

Figure 12: Comparison of distributions of proportion of highly correlated pairs on chromosome 3 between Aorta dataset and CATHGEN dataset. From the Aorta dataset, the distribution showed a peak in 115Mb-135Mb indicated by red arrow, but there was no peak at the same position from distribution come from the CATHGEN dataset. 52

Figure 13: Comparison of distributions of proportion of highly correlated pairs on chromosome 3 between the Aorta dataset and the GDS1597 dataset. There was no peak at the same position in the GDS1597 dataset. 54

Figure 14 Complementary cumulative distribution functions of R_S. The figure shows variation of R_S by changing $\rho_{threshold}$. The black shadow demonstrates the distribution of R_{S1} to R_{SN}. The red line demonstrates proportion of highly correlated pairs between tags in 115Mb-135Mb on chromosome 3, *S*, and the others in the genome, *S'*. (a) The distribution displays the results from the analysis on the Aorta data set, and (b) shows the results from CATHGEN data set. (c) Distribution of proportion of high correlated pairs from randomly sampled tags. This figure comes

from dissection of (a) at |ϱ|=0.7. The false discovery rate is 0.47%. While analyzing the Aorta dataset, permutation number (N) is set at 3,000; while analyzing the CATHGEN dataset, N is set at 100. 56

Figure 15: Connections between genes involved in highly correlated pairs more than 10 times. The red edges represent positive correlation coefficients, and the blue edges represent negative correlation coefficients. 65

Figure 16: Copy-number variations reported in 115Mb-135Mb on chromosome 3. The red blocks indicate reported gain segments, blue blocks indicate loss segments, and green blocks indicate variations (could be gain or loss). 66

Acknowledgements

First and foremost, I wish to thank to my advisor, Dr. Elizabeth Hauser, for inviting me to join her project. It is surely a real honor to explore the human genetics with one of the most innovative mind in the field. I would like to thank Dr. Sayan Mukherjee and Dr. Terry Furey for giving critical opinions on statistical models. I would also like to thank Dr. Williams Kraus, Dr. Simon Gregory, and Dr. Svati Shah for tutoring me the biological concepts related to cardiovascular disease. I would like to thank Dr. Jeanette McCarthy, the DGS of my program, who helped me in many ways while I participated in the program. I would like to thank all members of METHOD and CVD groups in the Center of Human Genetics. They gave me graceful advises and enlighten possible exploration in my thesis project. My work depended heavily on the numerous resources from the Center of Human Genetics at Duke University Medical Center and the Institute of Genome Science & Policy. None of my work would have been possible without generous funding from my program – Computation Biology & Bioinformatics – and Duke Graduate School. Finally, I would like to thank my parents and my friends. Without their selfless supports, I would not possible to finish my degree.

1. Introduction

1.1 Cardiovascular Disease

Cardiovascular disease (CVD) is the leading cause of disability and death in the industrialized world [1]. Classic epidemiologic studies have identified many personal and environmental risk factors for CVD, including age, sex, hypertension, dyslipidemia, diabetes mellitus, smoking, and physical inactivity [2].

One category of the cardiovascular disease is occlusive vascular disease, which is defined by the irregular accumulation of vascular smooth cells, inflammatory cells, and extracellular matrix proteins within the space between the endothelial lining and the media layer. One of the most common forms of occlusive vascular disease is atherosclerosis, a condition in which artery walls thicken [1].

Gene expression is a biosynthesis process that uses genetic messages to create functional genetic products. These genetic products contain proteins and functional RNA such as rRNA and tRNA. All known types of life – viruses, prokaryotes, and eukaryotes – have gene expression to operate all kinds of functions. The information carried by genetic codes – the genotype – is interpreted by gene expression, and the outcomes of expression give rise to phenotypes.

The original form of a gene – i.e. a long chain of DNA – is not active. The stretch of DNA is just the blueprint for the whole process. A process called transcription turns DNA into RNA. Transcription is performed by RNA polymerase, which builds RNA

strands by adding one RNA nucleotide at a time. The assembled RNA strands are complemented to the original DNA – for example, a C on the DNA will transcript a G on RNA.

In eukaryotic cells, primary transcript RNA would undergo post-transcriptional modification to convert into mature RNA. The most common post-transcriptional modification is RNA splicing. During RNA splicing, non-coding regions called introns are removed and coding sequences, exons, are joined.

Gene expression is delicately controlled and regulated to ensure phenotypes are presented in the right places at the correct time [3]. By ensuring the correct transcript of RNA and translated proteins, genes can express its designed roles and present destined functions to support life.

However, there are events that DNA itself changes. DNA damage happens when DNA is under potential attacks spontaneously or from external sources. Some types of the external factors causing exogenous DNA damage are (a) ultraviolet radiation, (b) ionized radiation, (c) alkylating agents, (d) hydrolysis agents, and (e) viruses [4, 5]. On the other hand, the endogenous DNA damage is the unavoidable outcome of a large complex molecule present in a metabolically active environment. The examples of spontaneous mutagenesis include depurination, the interaction between oxygen and free radicals, and errors caused by DNA replication [4].

Through the evolution process, variations in gene expression are preserved in different populations. Single-nucleotide polymorphisms (SNPs) are one of the results of evolution with mutations. A SNP is a DNA sequence variation of a single nucleotide in the genome differing between members of the same species, even with the same individual. For example, two DNA sequence fragments from different human, AACCGATC and AACCTATC, contain a mismatch in a single nucleotide. Mostly, common SNPs have only two alleles. In the previous case we say that the SNP has two alleles: G and T. During translation, the protein sequences coded from DNA sequences might change due to SNP variations. The alternations of phenotypes might cause malfunctions in cell processes and finally induce genetic disorders and diseases. Recently, many studies have demonstrated that SNPs are associated with atherosclerosis, and detection of SNPs becomes a powerful tool to discover the genes suspected of causing cardiovascular disease [6-10].

Another reason which might cause gene expression variation is copy number variation (CNV). A CNV is a segment of DNA in which copy-number differences have been detected by comparison of two or more genomes. The length of segment could range from one kilo-basepairs to mega-basepairs [11]. Ordinarily, being diploid, humans have two copies of each autosomal region, one per chromosome. However, the copy-number may vary for specific regions resulting from duplication or deletions. CNVs can be limited to a single gene or include a contiguous set of genes. Differing by duplications

or deletions, CNVs can result in having either too many or too few of the quantity-sensitive genes, which may be responsible for the variability of human phenotypes [12, 13]. Previous research has shown that gene variations caused by CNVs could be one of the reasons causing cardiovascular diseases [14-19].

As with other complex human diseases, the etiology of atherosclerosis is likely due to interactions between multiple genetic and environmental factors [20]. Genetic predisposition to atherosclerosis consists of multiple and relative genetic variants with small effects. The genes can manipulate risk of disease alone or cooperated with environmental factors. [21-24]. There are many different kinds of environmental factors. First, they could be from outside the body. For example, smoking, diet, and stress have been known to be risk factors of cardiovascular disease. Furthermore, the environmental factors could be expressed as tissue-specific expressions in the body. Some genes could be highly expressed in one region and rarely expressed in another region at the same time. By identifying genes expressed variously in different locations, we could infer the tissue-specific factors which alter the gene expressions locally and result in disease phenotypes [25-27].

In addition to tissue-specific factors, the genes associated with cardiovascular diseases are also located in several suspected regions in the genome. Previous genome-wide linkage studies analyzed genetic linkage and showed that regions on chromosomes 1q25, 3q13, 5q31, 7p14, and 19p13 are highly associated with coronary artery disease

[28]. In these regions, chromosome 3q13 was shown to have the most significant multiple logarithm of odds score (LOD score) among the genome. Furthermore, some genes located in or near the chromosome 3q13 region have been shown to be associated with cardiovascular disease. These genes include GATA2 [29, 30], LSAMP [31], and KALRN [2, 30, 32]. There is also some previous research showing that chromosome 3q in genome is a region suspected to have a strong relationship to cardiovascular disease [28, 33-35].

1.2 Gene Expression and Correlation Coefficients

Modern biotechnologies, unlike single gene experiments, allow researchers to simultaneously monitor large numbers of gene expression. For example, researchers are capable of investigating numerous gene expression levels and examining the regulation and the interactions between genes by DNA microarray experiments. To perform gene expression analysis, expressed RNA fragments which degrade rapidly are transferred into complementary DNA (cDNA) after a reverse transcription process to fix expression level. After reverse transcription, cDNA are hybridized to microarray system, and gene expression levels are quantified by measuring florescence intensities.

Microarray data analysis focuses on comparing differentially expressed genes in control and in treatment groups and deriving the dependency relationships in sets of differentially expressed genes. One way to achieve this goal is to perform two-color microarrays. Two-color microarrays label cDNA samples with two different

fluorophores, Cy3 and Cy5. The Cy-labeled samples are mixed and hybridized on single microarray [36]. Researchers compare gene expression levels between control and treatment groups by measuring differences of florescence intensities from Cy3 and Cy5. The ratio of intensities is used to cluster genes into groups and the dependency structure of differentially expressed genes can be examined. The other method to examine gene expression levels is measuring absolute florescence intensities from a single-channel microarray. In single-channel microarrays or one-color microarrays, the arrays provide intensity data for each probe or probe set indicating a relative level of hybridization with the labeled target. One of the benefits of single-channel microarrays is that an aberrant sample would not affect the raw data derived from other samples because each array chip is only exposed to one sample and hybridized once. Another benefit is that it is relatively easy to compare gene expression levels between samples. By adjusting the batch effects and normalizing, gene expressions from different samples and even experiments can be compared.

One of the methodologies to analyze dependency of expressed genes is to identify diverse correlation structures in gene expression patterns. By analyzing the correlations, it is possible to compare genes expression levels between individuals and determine genes that might be associated with disease [37-41].

1.3 Specific Aims

From previous studies, various genome wide correlation analysis methodologies have been applied on microarray analysis. However, none of these studies were performed in localized genomic region. With gene-oriented correlations tests, the scale of methodology is limited to individual genes. The methodology is relatively weak in identifying relationships of multiple factors associated with disorders. On the other hand, genome-wide studies including genome-wide association studies and genome-wide linkage studies may contain too much noise. Genome-wide studies in correlations included all data from the genome, which might filter out the subtle signals associated with disease. This study sought to balance between gene-oriented correlations and genetic associations by dividing the entire genome into regions. With the region-oriented correlation analysis, we were able to detect the regions possibly associated with cardiovascular disease as those regions having more significantly highly correlated gene expression pairs with empirical P value < 0.05. Then, we looked into the regions and selected genes which are involved in high correlation. These detected genes could be validated by other databases and previous research, showing that they might have associations with cardiovascular diseases.

2. Materials and Methods

2.1 Subjects

Seven independent expression data sets were analyzed in this study. Two data sets were generated at Duke University Medical Center and the others were obtained from Gene Expression Omnibus database (GEO). The clinical characteristics of these data sets are shown in Table 1.

Duke University Aorta data set (Aorta) – 122 samples from human aortas were collected at organ harvest from 78 heart transplant donors. Each aorta section was graded for atherosclerosis burden by two methods. Sudan IV, which stains lipids and low-density lipoprotein (LDL), assesses early atherosclerotic plaques with imaging processing software [20]. The advanced atherosclerotic degree was determined by the amount of raised lesions. Raised lesions are designated as the combinations of fibrous plaques, complicated lesions, and calcified lesions [42]. RNA samples are extracted from aorta sections as described [20]. RNA samples from the aorta data set were analyzed on Affymetrix GeneChip® Human Genome U95Av2 Array platform. Affymetrix Genechip® Human Genome U95Av2 Array has 12625 probes on the chip.

CATHGEN data set – This dataset is a subset of the larger CATHGEN biorepository. Patients undergoing catheterization were recruited through the cardiac catheterization laboratories at Duke University Hospital (Durham, NC, USA) with approval from the Duke Institutional Review Board. Medical history and clinical data

Table 1: Clinical characteristics of the seven expression data sets.

Characteristic	Data set						
	Aorta	CATHGEN	GDS1597	GSE2014	GSE13985	GSE6088	GDS268
Total number of individuals	78	191	22	24	10	23	24
Number of RNA Samples	78	191	51	25	10	23	24
Age (in years)							
Mean	36.86	58.43	46.3	48.63	34.1	N/A	N/A
Std	15.86	13.03	7.8	13.46	7.94	N/A	N/A
Race (%)							
Caucasians	87.18	66.49	59	N/A	N/A	N/A	N/A
Africa American	10.26	27.23	14	N/A	N/A	N/A	N/A
Sex (%)							
Male	61.54	59.16	77	83	100	N/A	0

were collected and stored in the Duke Databank for Cardiovascular Disease maintained at the Duke Clinical Research Institute [2, 43]. RNA samples are obtained from white blood cells collected in special tube designed to stabilize the RNA. For this analysis, we selected 191 samples with sufficient quantity and high quality of RNA samples. The RNA samples from the CATHGEN data set were analyzed on the Affymetrix GeneChip® Human Genome U133 Plus 2.0 Array with 54675 probes.

GDS1597 data set – Major epicardial coronary arteries were dissected from the explanted hearts of patients undergoing orthotopic heart transplantation. 55 RNA samples were extracted from 21 human arteries and were analyzed by Agilent TQ Cardiovascular Human Oligonucleotide Array [44]. This Agilent microarray is spotted with 20433 probes on the chip. Since there was more than one sample for each individual, we randomly chose one gene expression profile from each individual to prevent over-estimating correlations.

GSE2014 data set – 25 myocardial samples were obtained from the left ventricle free wall from the patients undergoing heart transplantations. The RNA samples were analyzed on a customized 7744-probe cDNA array from the Norwegian Microarray Consortium [45].

GSE13985 data set – Ten RNA samples were taken from all white blood cells GSE13985 from Slovenian patients diagnosed with Familial hypercholesterolemia, which is an autosomal dominant disorder caused by mutations mainly located in the low-

density lipoprotein receptor (LDLR) gene [46]. The RNA samples were analyzed on the Affymetrix GeneChip® Human Genome U133 Plus 2.0 Array.

GSE6088 data set – 23 RNA samples were taken from CD3 positive T cells from German patients diagnosed with Familial hypercholesterolemia. The RNA samples were also analyzed on the Affymetrix GeneChip® Human Genome U133 Plus 2.0 Array.

GDS268 data set – Skeletal muscle cells were sampled from patients undergoing elective abdominal surgery. These patients are in different stages of obesity. The extracted RNA samples were analyzed on Affymetrix GeneChip® Human Genome U133A Array with 22283 probes on the chip [47].

The raw intensity data from all microarray systems were converted to a log2 scale followed by a quantile normalization [48]. After normalization, the seven data sets were analyzed separately using the statistical methods described below. We had access to the normalized data for all seven data sets.

2.2 Statistical Analysis

2.2.1 The Localized Pearson Correlation Test

Our goal was to analyze the expression correlations in the region 115Mb-135Mb on chromosome 3 defined by the linkage peak in the previous study [2, 28]. We were interested in evaluating correlation patterns in this specific genomic region and to compare these regions to the rest of the genome. We evaluated the connectivity of a region by correlation between pairs of expression tags. The Pearson's correlation was

computed among all microarray tags in the specific region. For example, if there are n microarray tags in the specific region, the Pearson's correlations of $C_2^n = n(n-1)/2$ possible pairs would be computed. Next, the pairs of tags were selected if their correlation coefficient was above the threshold to define highly correlated pairs. We varied the correlation threshold from 0.5 to 0.8. Afterward, we computed the proportion of highly correlated pairs:

$$R = \Pr\left(\#\left\{|\rho_{ij}| \geq \rho_{threshold}\right\}\right)$$

$$= \frac{\#\left\{|\rho_{ij}| \geq \rho_{threshold}\right\}}{\#\left\{\rho_{ij}\right\}}$$

where $1 \leq i < j \leq n$. In most of the analysis performed, the threshold of the absolute value of correlation coefficient was set at 0.7.

We were interesting in explaining the expression of genes in the region of chromosome 3 defined by the linkage peak as a fixed genomic region as well as taking the disease associated genes and the differentially expressed genes into account. In this study, we compared two different methods to define genomic regions of interest. One method was to evaluate all tags in windows with fixed length of basepairs to match the linkage region. The whole genome was divided into 20Mb windows with step size 1Mb. Then, the Pearson's correlations were computed in every window. This bp-size-based window definition is shown as Figure 1 (a).

(a)

(b)

Figure 1: Definition of regions relative to expression tags. The red segments in figures show tags on a microarray. One strategy is that the size of the sampling region and the step size are based on a fixed length in basepairs. For example, in (a), the length of the sampling region is fixed at 20Mb. The other strategy is that the size of the sampling region and step size are based on a fixed number of tags. In (b), the length of the sampling region is fixed at 4 tags, for example.

The other method was tag-oriented. By this method, the number of tags in every window was fixed to a certain number. In the study, we set the number of tags close to number of tags in 115-135Mb on chromosome 3. The number of tags varied from platform to platform depending on how many tags which were on the chip are in the region of interest. For example, there are 95 tags on the Affymetrix U95Av2 system used by the Aorta study. Therefore, the sampling window for the Aorta dataset was set at 95-tag long. This tag-based window definition is shown as Figure 1 (b).

After determining the length of window, we expanded the correlation analysis to the whole genome. The high correlation proportion, *R*, was obtained and the distribution of *R* from each window was plotted. From the whole genome screening

process, the regions with significant high R (P value < 0.05 by Z-test) were selected for further analysis. Furthermore, the distributions of R between different datasets were compared.

We then tabulated the frequency of genes appearing in the high correlation list. The relationships between these pairs were checked, including the physical distances, functional relationships, and genetic association results.

2.2.2 Permutation Test

We are interested in how the linkage region of 115Mb to 135Mb on chromosome 3 compared to the rest of the genome. It was too computationally intensive to examine all possible pairwise correlations in whole genome, so we performed random sampling in the genome. Our goal was to generate a distribution of the proportion of correlation between tags under the null hypothesis of no regional correlation. Let us denote the set of microarray tags in chr3:115-135Mb as S. Relative to other regions, S exhibits a high correlation proportion. To evaluate this high level of correlation proportion, we computed the correlations pair-wisely between S and S', which is the complement of S. Next, the percentage of high correlation pairs from all possible pairs between two sets S and S' was computed, denoted as R_S. Next, the 95 microarray tags denoted as S_1 were randomly selected from the whole genome, then, the set of correlation coefficients of all possible pairs between two sets S_1 and S_1' was computed, denoted as $\{\rho_{S_1 S_1'}\}$. The

proportion of high correlation pairs, R_{S1}, was also calculated. The random sampling process and analysis procedure were repeated by N times and R_{S1} to R_{SN} were obtained through the computations. Finally, FDR was computed from the result of permutations:

$$FDR = \Pr(R_{Sn} \geq R_S)$$

where $1 \leq n \leq N$ [49, 50]. For the Aorta dataset, we drew 3,000 random samples (N = 3000); and for the CATHGEN dataset, we drew 100 random samples (N=100).

Perl and R on the Solaris system were used for all statistical analysis.

3. Results

3.1 Pair-wise correlation between 115Mb and 135Mb on chromosome 3

The correlation analysis was applied to all seven datasets in the region of 115Mb to 135Mb on chromosome 3. Table 2 shows the analysis results. For the Aorta dataset using the Affymetrix U95Av2 microarray platform, 95 tags were extracted from the region between 115Mb and 135Mb on chromosome 3. Among 4465 possible pairs, we found that there were 348 (7.79%) pairs whose absolute values of correlations were over $\rho_{threshold} = 0.7$.

The same analysis was applied on the CATHGEN data set. However, the analysis showed only 338 (0.30%) pairs were kept after selection by $\rho_{threshold}$ from 113050 possible pairs constructed by 476 tags between 115Mb and 135Mb on chromosome 3 from Affymetrix U133 plus 2 microarray platform.

The analysis was also applied to the five data sets downloaded from GEO. After the analysis, GDS1597 and GSE13985 were also observed having similarly large proportion of highly correlated pairs. For the region 115Mb to 135Mb on chromosome 3, *R* of GDS1597 was 12.27% and *R* of GSE13985 was 4.27%. The proportions of highly correlated pairs from GSE2014, GSE6088, and GDS268 were quite low. *R* values from these three data sets were 0.455%, 0.131%, and 0.305%, respectively.

Table 2: Proportion of the pairs with high correlation in the 115Mb-135Mb region on chromosome 3 compared across seven expression data sets.

	Data set						
	Aorta	CATHGEN	GDS1597	GSE2014	GSE13985	GSE6088	GDS268
Total number of tags in the region	95	476	99	56	476	476	234
Possible number of pairs	4465	113050	4851	1540	113050	113050	27261
Number of pairs with $\lvert\rho\rvert \geq 0.7$	348	338	595	7	4829	148	83
Proportion of pairs with $\lvert\rho\rvert \geq 0.7$ (%)	7.79	0.299	12.27	0.455	4.27	0.131	0.304

3.2 Characteristics of highly correlated genes from the Aorta dataset

After analyzing 115Mb to 135Mb on chromosome 3 of the Aorta dataset, we observed 348 highly correlated pairs with $|\rho| \geq \rho_{threshold} = 0.7$. From the 348 pairs, the distances between the highly correlated pairs were analyzed. Figure 2 demonstrates the distribution of distances between highly correlated pairs. The range of distances was

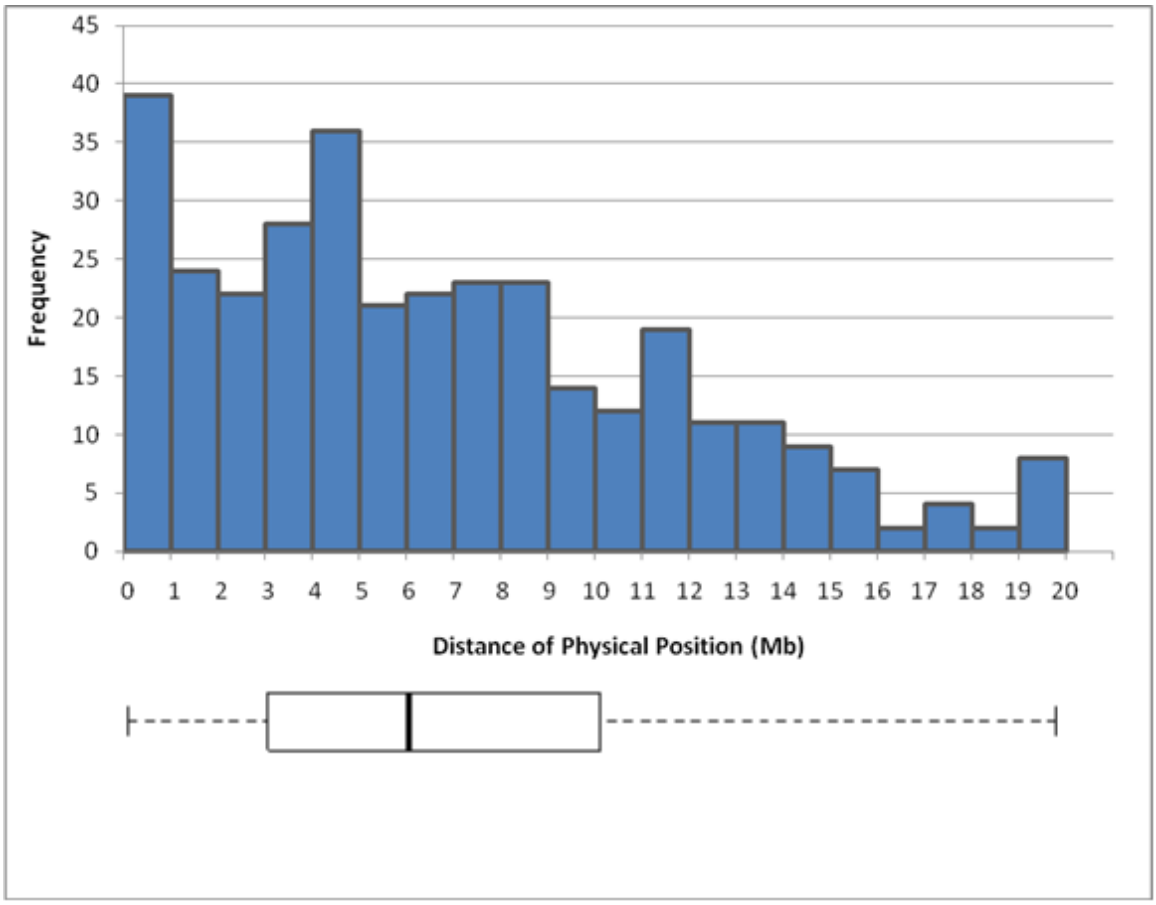


Figure 2: Distribution of distances of physical position between highly correlated pairs in 115Mb to 135Mb on chromosome 3 from the Aorta dataset. The bar plot shows the frequency of distances between highly correlated pairs. The lower panel, the box plot, shows the range, quartile, and median of distances of the highly correlated pairs.

between 129bp and 19.8Mbp, and the median of distances is 5.97Mbp. The highly correlated pairs did not appear to cluster in any smaller region within the 20Mb window and could be as far as 20 Mb apart.

These 348 highly-correlated pairs were constructed by 53 Affymetrix tags in the region. With further matching, these 53 Affymetrix tags are grouped into 45 genes in the region. These 45 genes and the frequencies of appearances are displayed in Figure 3. By applying $\rho_{threshold} = 0.7$, SLC15A2 and RHO, with 43 appearances in the highly correlated pairs, were the gene which appeared most in the highly correlated pairs.

The genes previously found associated with cardiovascular disease, *e. g.* GATA2 and LSAMP, were among the genes involved in highly correlated pairs.

3.3 GATA2 and LSAMP are at the center of the connections

After the statistical analysis in the region 115Mb to 135Mb on chromosome 3 from the Aorta dataset, we gathered information about the genes associated with highly correlated pairs. From the analysis, 26 genes were found in the list of highly correlated genes with greater than or equal to ten connected neighbor genes. Some genes in the set, such as GATA2, LSAMP and RHO, have been reported as associated with atherosclerosis [29, 31]. According to our analysis, GATA2 and LSAMP not only showed high correlation to each other with $\rho = 0.7846$ (Figure 4), they also tended to connect to each other tightly by having high and *cis*-directed correlation with the genes

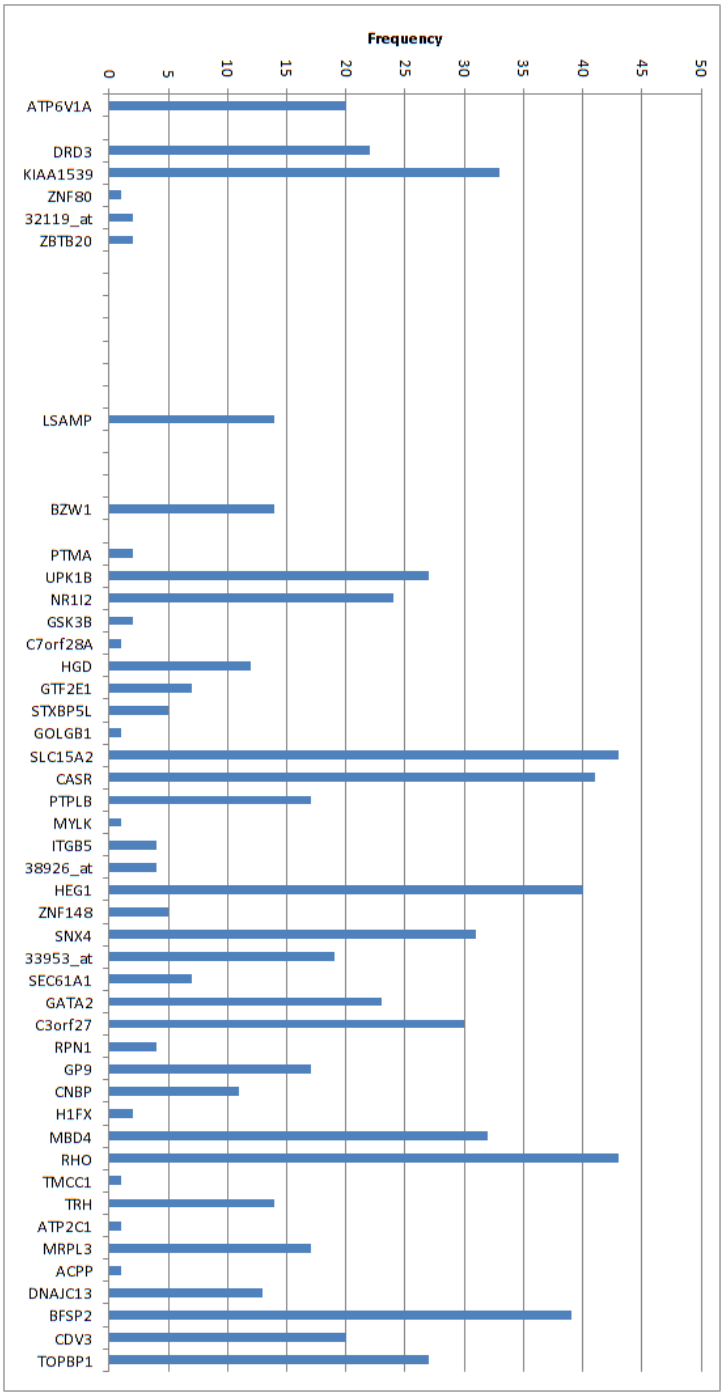


Figure 3: The frequencies of genes appearing in the high correlation pairs in 115Mb to 135Mb on chromosome 3 from the Aorta dataset. The genes are arranged by the order of physical locations.

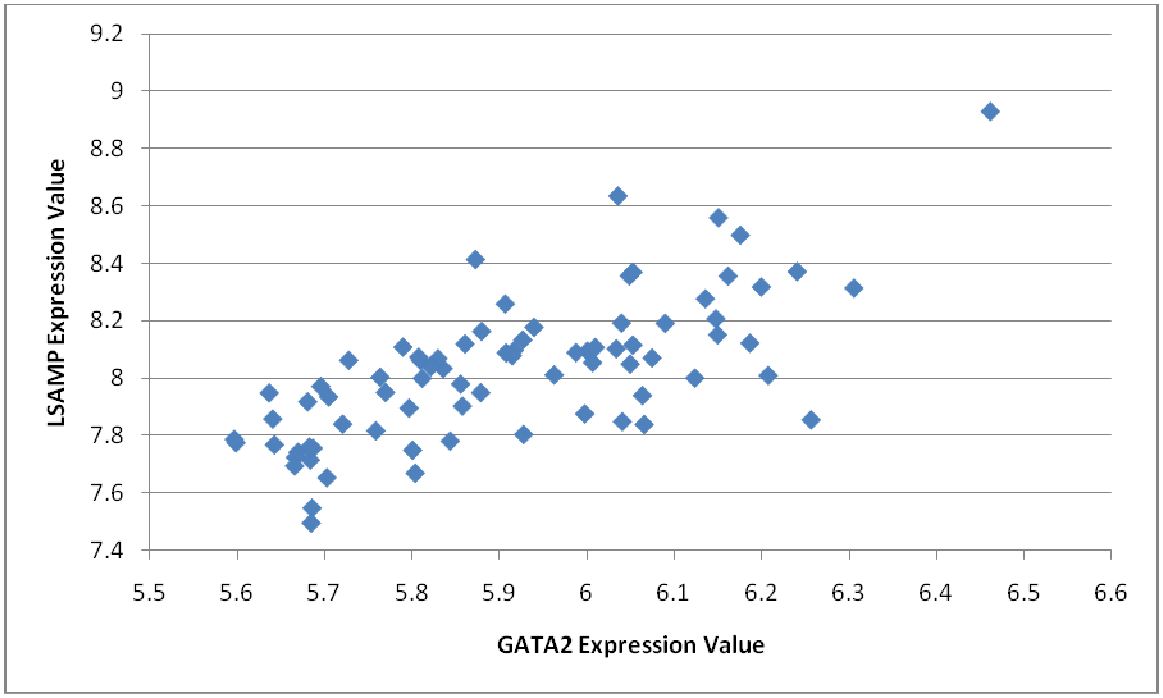


Figure 4: Scatter plots of normalized gene expression values of GATA2 and LSAMP in the Aorta dataset. The correlation coefficient between GATA2 and LSAMP expressions is 0.7846.

they are connected to (Figure 5). For example, GATA2 and LSAMP both showed highly positive correlation with DRD3 and highly negative correlation with ATP6V1A. If multiple Affymetrix tags are included in one genes, all of these tags also tended to show *cis*-directed correlations, with the exception of HEG1. However, analysis of GATA2 and LSAMP in the CATHGEN dataset showed no evidence of connections between these genes (Figure 6).

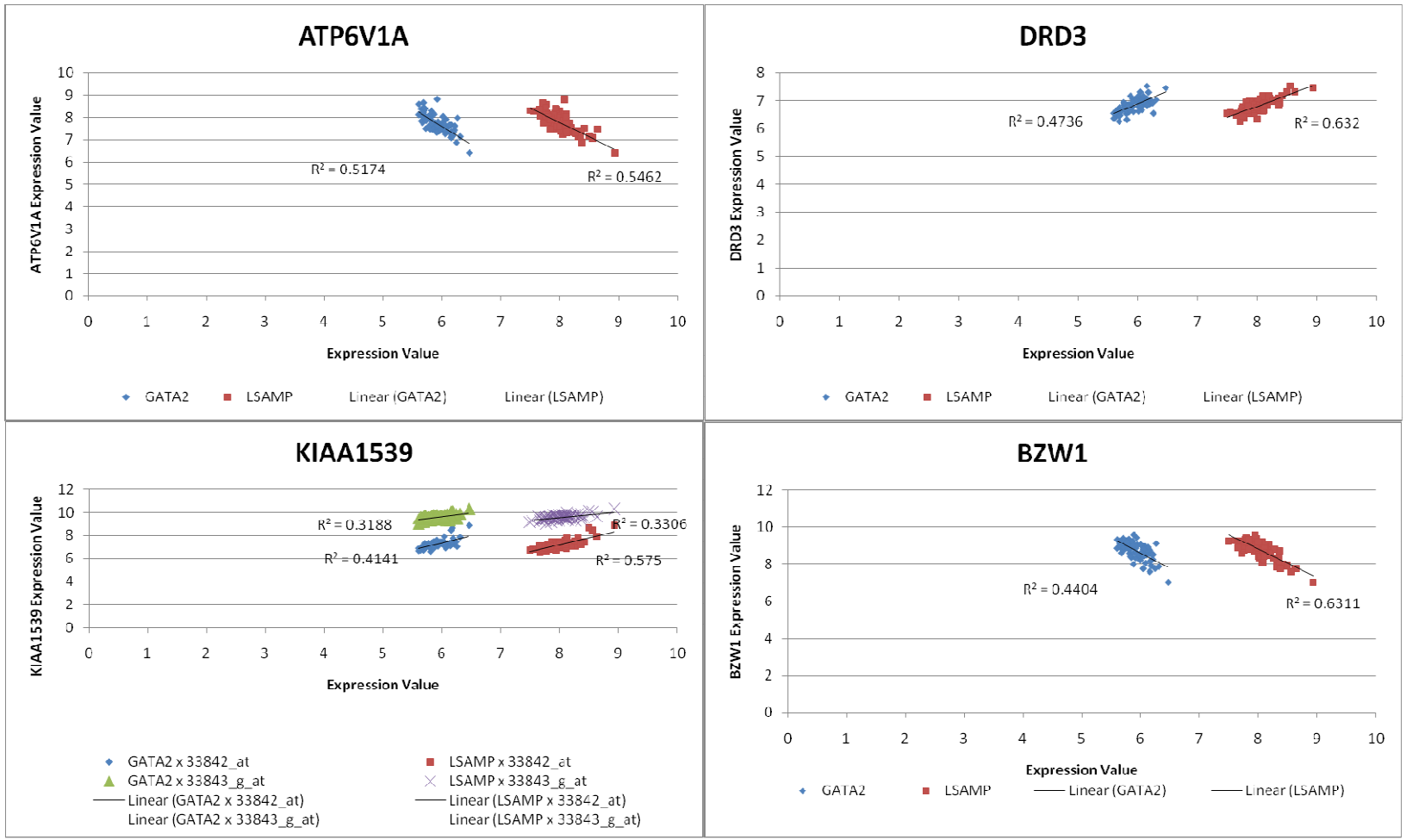


Figure 5: Scatter plots of GATA2 and LSAMP with the genes that are highly correlated to both GATA2 and LSAMP in the Aorta data set. The correlations demonstrate a *cis*-directed property, except for HEG1.

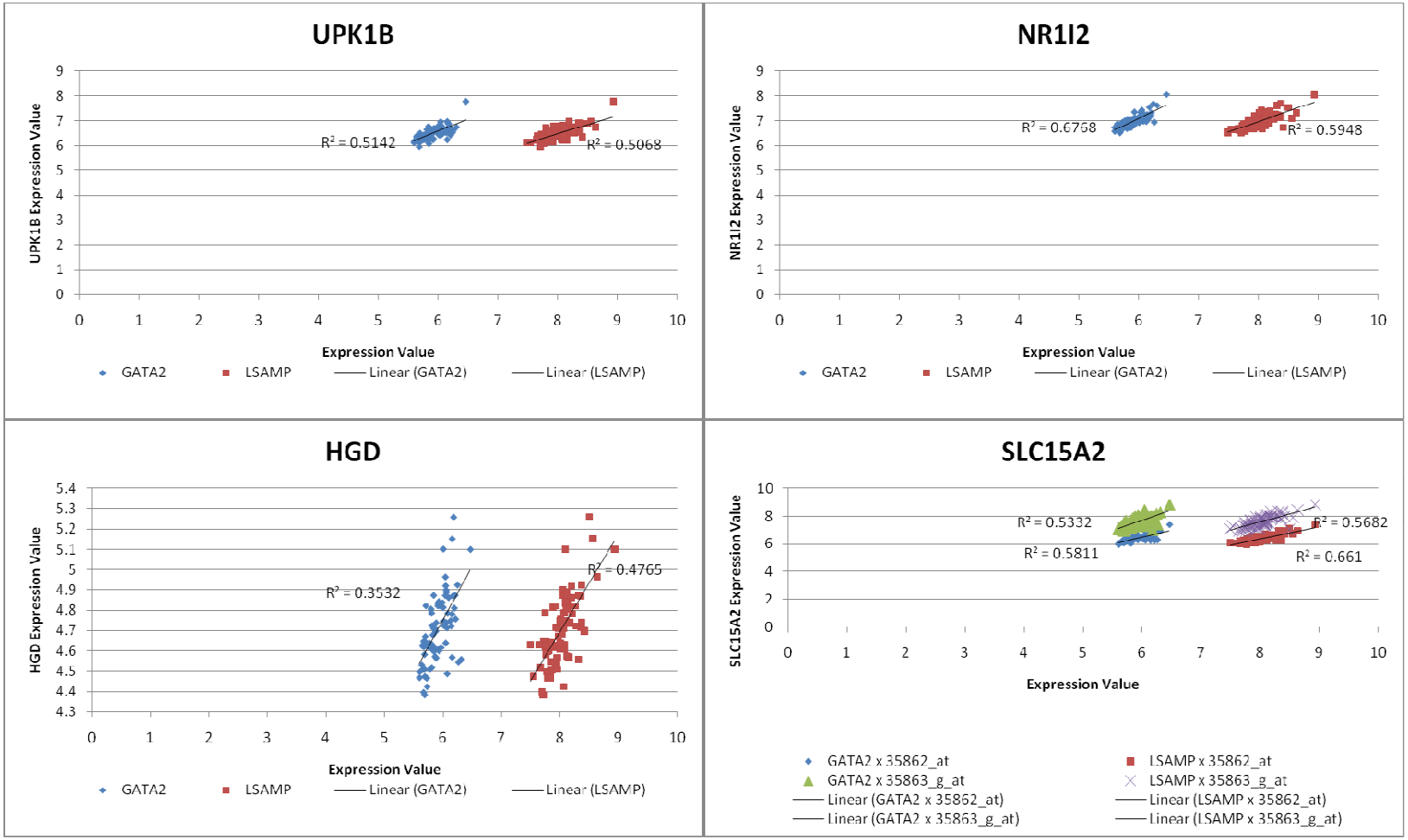


Figure 5: Cont'd.

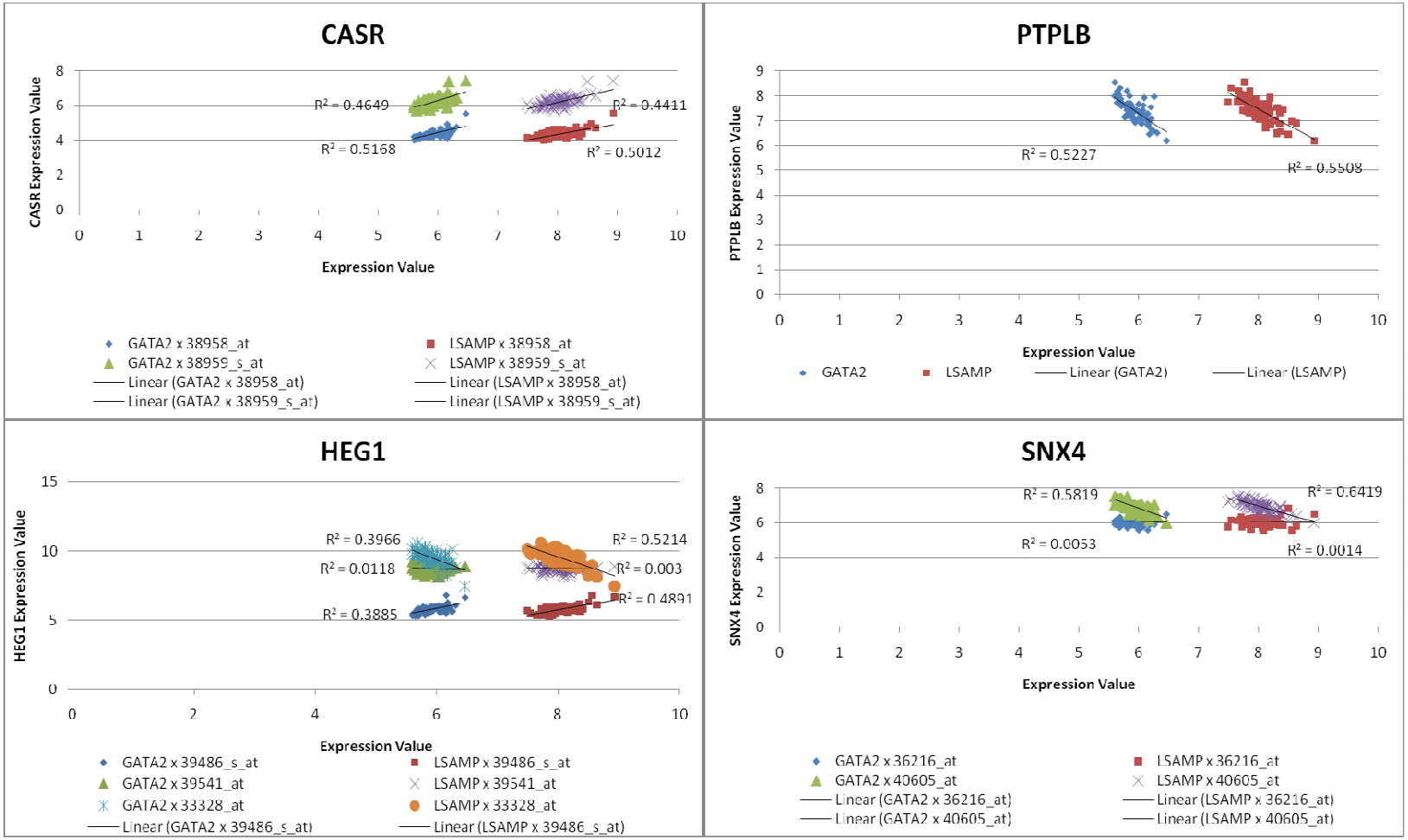


Figure 5: Cont'd.

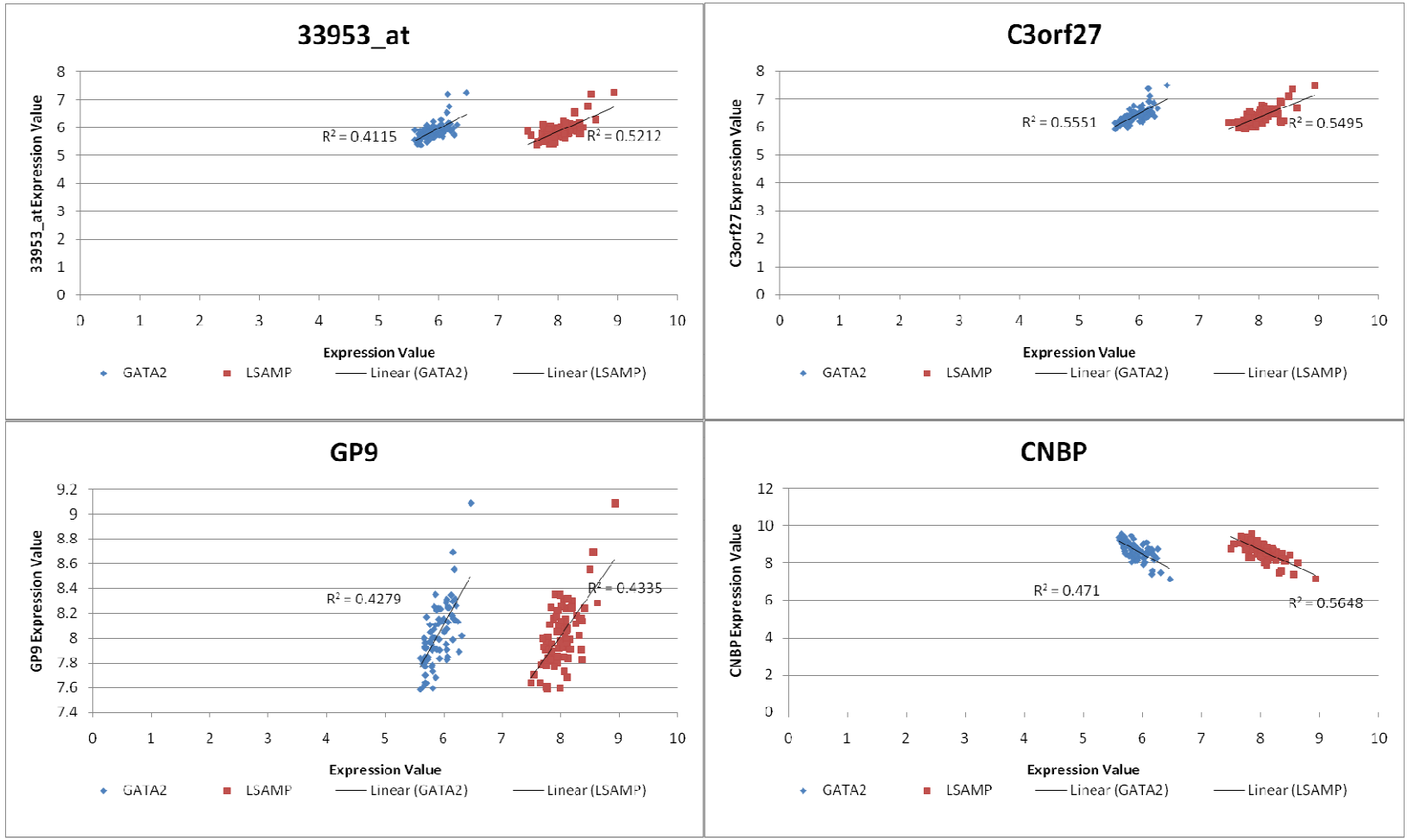


Figure 5: Cont'd.

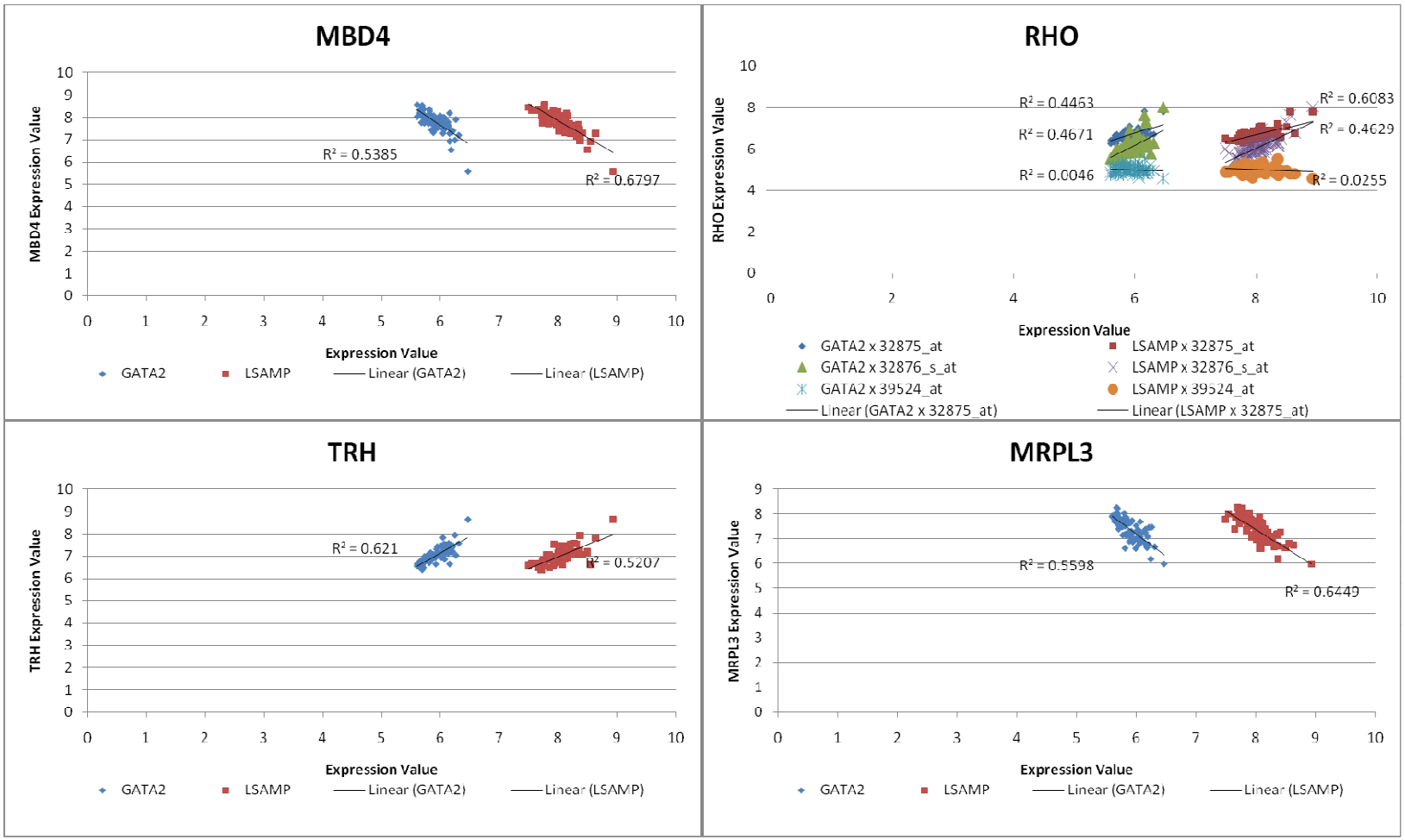


Figure 5: Cont'd.

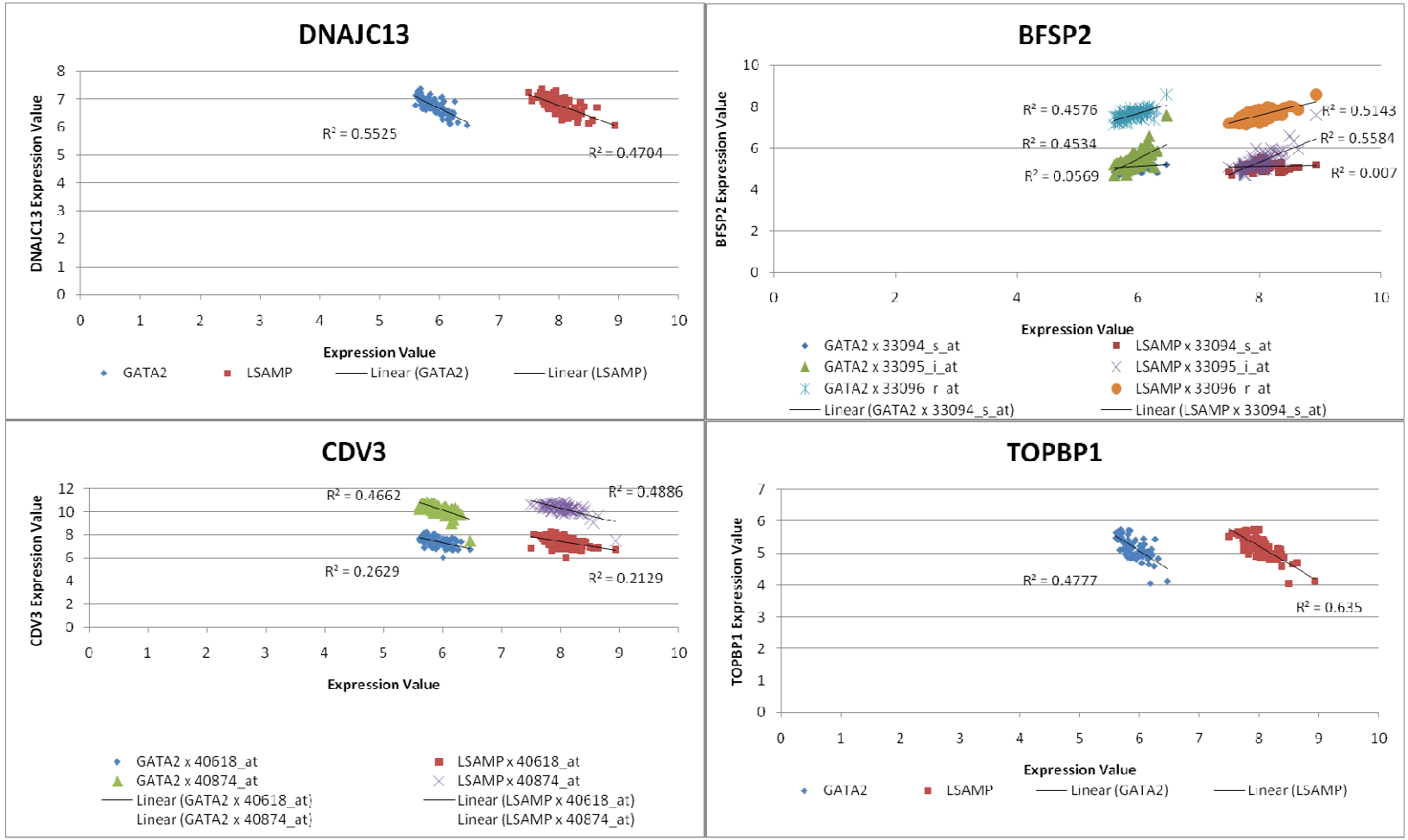


Figure 5: Cont'd.

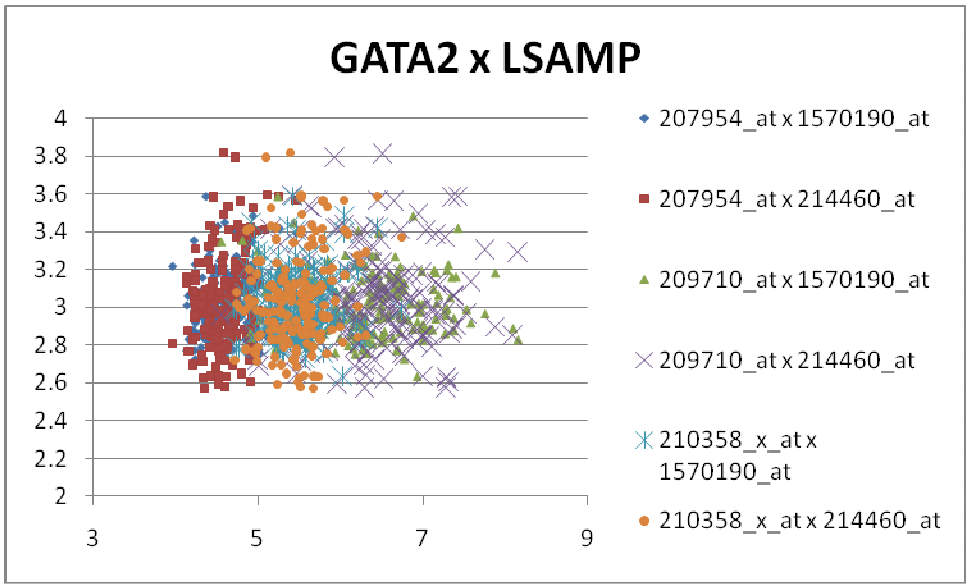


Figure 6: Scatter plot of GATA2 and LSAMP in the CATHGEN dataset. In U133plus2 microarray platform, Affymetrix IDs of GATA2 are 207954_at, 209710_at, and 210358_x_at, and Affymetrix IDs of LSAMP are 1570190_at and 214460_at. The scatter plot suggested that there is no relationship between GATA2 and LSAMP in the CATHGEN dataset.

3.4 115Mb – 135Mb on chromosome 3 is a region with significantly strong connectivity in whole genome

In order to understand the general distribution of the high correlation proportion among the genome, the correlation analysis procedures were expanded to the entire genome. Divided into 20Mb windows, the strength of connection in each region was estimated by calculating the proportion of highly correlated pairs, *R*, as described in the methods.

In the analysis of the Aorta data set, the genomic sequences were separated into 2533 sampling windows. Among the sampling windows, the mean of the proportion of high correlated pairs is 4.04% and the standard deviation is 3.62%. From the result, the

percentile of *R* in 115Mb-135Mb on chromosome 3 is 96.76%, which means this region is strongly connected within the whole genome. The cumulative density function is shown as Figure 7.

One difficulty in analyzing windows of fixed size, say 20Mb, is that the cDNA tags on the Affymetrix HG-U95Av2 microarray do not distribute uniformly through the genome. Some regions, *e. g.* the centromeres, have fewer tags. A region with a small number of tags results in a small denominator in calculating *R*. With a small denominator, the window might have extraordinary high *R*, and the extreme *R* values would shift the distribution of *R* away from the normal distribution as Figure 8 (a).

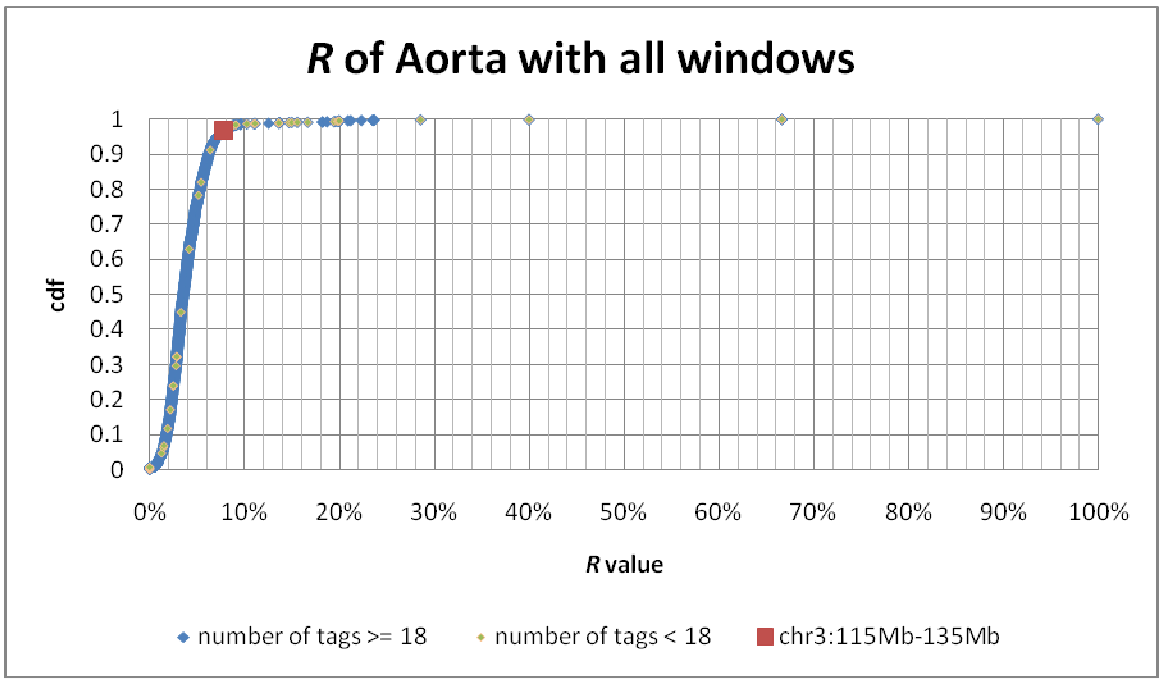


Figure 7: The cumulative distribution function of proportion of high correlation coverage, *R*, from the Aorta data set. The red dot represents value of chr3:115Mb-135Mb. The sampling windows included all possible *R* values without applying a filter for the number of tags.

If we included all windows in the distribution, although the percentile of proportion of highly correlated pairs in chr3:115Mb-135Mb is 96.76%, the *P* value from a Z-test of *R* value in chr3:115Mb-135Mb is 0.1504.

In order to adjust for a small number of tags in a window of fixed size, we applied a cutoff value for number of tags in a window. From the Aorta dataset, the number of tags in the region of the centromere of chromosome 1 is less than 18. If we set the threshold of number of tags at 18, the *R* value would not be included in the distribution. Sixty-seven of 2,533 (2.65%) sampling windows were eliminated after applying the filter for the number of tags. The mean of *R* became 3.87% and the standard deviation shrinks to 2.03%. Therefore, the *P* value of R in the region 115Mb-135Mb on chromosome 3 is 0.0266, which shows significant difference from other regions. Figure 8 (b) displays the distribution of *R* after filtering out regions with fewer than 18 tags.

Fixing the number of tags in the sampling window is another strategy to adjust for the uneven tag distribution screening the whole genome. Here we fixed the window size at 95 tags, the same as the number of tags in chr3:115Mb-135Mb. In this case, the entire genome is divided into 11484 windows. The mean of proportion of highly correlated pairs (*R*) was 3.96% and the standard deviation was 1.52%. Under this condition, the proportion of highly correlated pairs in chr3:115Mb-135Mb was still 7.79%, but the *P* value from a Z-test for the region was 0.0058, which strongly shows

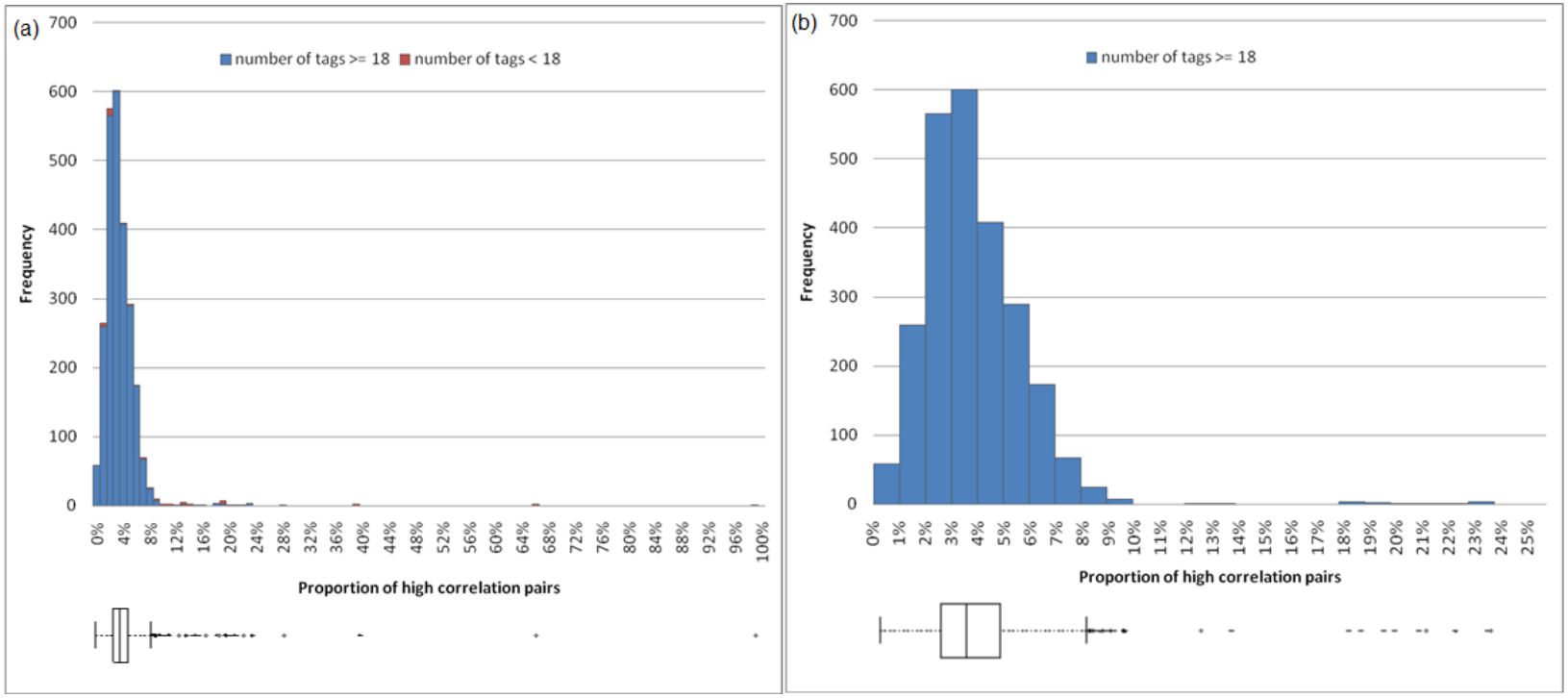


Figure 8: Distribution of proportion of high correlation coverage, *R*, from Aorta dataset. (a) The regions with less than 18 tags are included in the distribution and represent in red blocks. (b) The regions with less than 18 tags are filtered out. Many extreme values are eliminated and the distribution of *R* becomes closer to a normal distribution. (c) The distribution of *R* after applying tag-oriented sampling method. By fixing the number of tags in a window, the extreme values are eliminated.

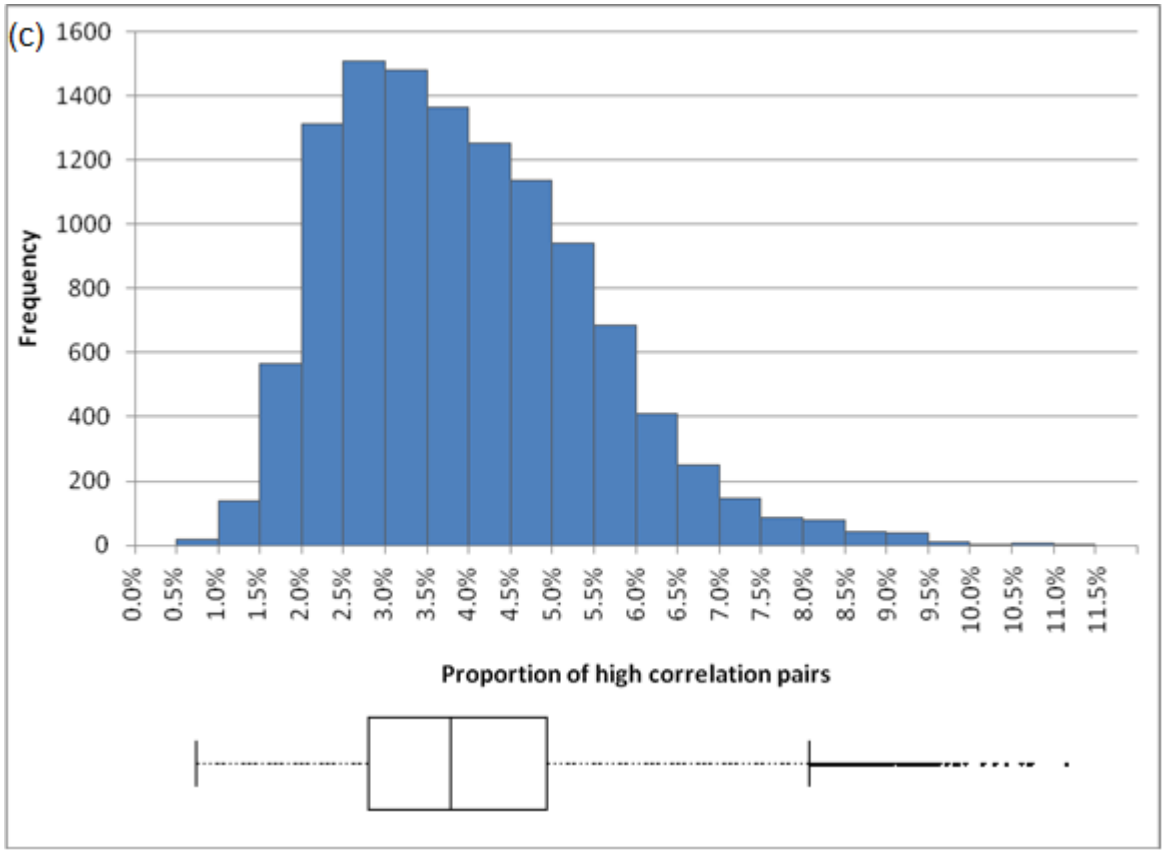


Figure 8: Cont'd.

a significant difference from other regions. Figure 8 (c) demonstrates the distribution of *R* while applying tag-oriented sampling method.

3.5 Some other regions in the genome showed significantly strong connectivity

After applying statistical analysis to the whole genome, we were able to discover the regions with extreme proportions of highly correlated pairs. While performing the screening process, the lower bound of the number of tags in a region was set at 18. According to the analysis on the Aorta dataset, Table 3 lists all regions which showed significantly strong connectivity (P value < 0.05) from the whole genome by applying a Z-test on the proportion of highly correlated pairs. In 2466 sampling windows which pass the threshold of number of tags, 94 sampling windows were detected having significantly large proportions of highly correlated pairs. Furthermore, these 94 sampling windows were not randomly dispersed in the genome. Here we merged windows into wider regions if the range of sampling windows overlapped at least 10Mb. After combining procedures, 94 sampling windows were grouped into 23 regions. The correlation analysis of the Aorta dataset for each chromosome is plotted in Figure 9.

Table 3: All regions with a significantly large proportion of highly correlated pairs (*R*) in the Aorta dataset. The sampling windows which overlapped with each other over 10 Mb were merged into a single region. The *P* value came from Z-test of proportion of highly correlated pairs.

Chr	Region (Mb)	Number of tags	highly correlated pairs		
			Number	Proportion	p-value
1	205-225	89	297	7.58%	0.0336
1	206-226	74	214	7.92%	0.0229
1	207-227	79	264	8.57%	0.0102
2	86-106	69	184	7.84%	0.0251
3	115-135	95	348	7.79%	0.0266
3	163-183	32	36	7.26%	0.0472
3	164-184	32	36	7.26%	0.0472
4	10-30	23	19	7.51%	0.0363
4	13-33	20	14	7.37%	0.0421
4	14-34	19	13	7.60%	0.0329
4	15-35	19	13	7.60%	0.0329
5	133-153	149	810	7.35%	0.0430
5	138-158	130	608	7.25%	0.0477
5	156-176	60	132	7.46%	0.0383
5	157-177	61	151	8.25%	0.0154
6	53-73	37	56	8.41%	0.0126
6	54-74	32	39	7.86%	0.0245
6	90-110	38	64	9.10%	0.0049
8	36-56	57	119	7.46%	0.0383
8	38-58	61	134	7.32%	0.0444
8	39-59	52	104	7.84%	0.0251
8	43-63	41	66	8.05%	0.0196
8	50-70	40	61	7.82%	0.0257
8	51-71	41	61	7.44%	0.0391
8	98-118	43	76	8.42%	0.0124
8	99-119	43	81	8.97%	0.0059
8	100-120	38	61	8.68%	0.0088
8	101-121	43	68	7.53%	0.0355
8	102-122	41	71	8.66%	0.0091
8	103-123	39	71	9.58%	0.0024
8	104-124	34	45	8.02%	0.0203
8	105-125	32	41	8.27%	0.0150
8	106-126	33	39	7.39%	0.0413

Table 3: Cont'd

Chr	Region (Mb)	Number of tags	highly correlated pairs		
			Number	Proportion	p-value
8	124-144	51	112	8.78%	0.0077
8	125-145	58	124	7.50%	0.0367
10	8-28	46	78	7.54%	0.0351
10	9-29	45	78	7.88%	0.0240
10	112-132	67	172	7.78%	0.0269
11	4-24	98	373	7.85%	0.0248
11	5-25	96	372	8.16%	0.0172
11	6-26	90	357	8.91%	0.0065
11	7-27	76	218	7.65%	0.0311
12	20-40	50	6	0.49%	0.0479
14	36-56	55	107	7.21%	0.0497
14	37-57	51	92	7.22%	0.0492
14	38-58	54	104	7.27%	0.0468
14	39-59	50	94	7.67%	0.0304
16	32-52	19	13	7.60%	0.0329
16	33-53	24	23	8.33%	0.0139
16	34-54	22	29	12.55%	0.0000
16	35-55	30	36	8.28%	0.0148
16	68-88	82	242	7.29%	0.0458
19	22-42	60	133	7.51%	0.0363
19	23-43	61	133	7.27%	0.0468
19	24-44	70	178	7.37%	0.0421
20	2-22	73	207	7.88%	0.0240
20	3-23	68	182	7.99%	0.0211
20	4-24	61	175	9.56%	0.0025
20	5-25	54	124	8.67%	0.0090
20	6-26	55	134	9.02%	0.0055
20	7-27	53	133	9.65%	0.0022
20	8-28	50	113	9.22%	0.0042
20	9-29	49	113	9.61%	0.0023
20	10-30	52	117	8.82%	0.0073
20	11-31	58	134	8.11%	0.0183
20	12-32	63	174	8.91%	0.0065

Table 3: Cont'd

Chr	Region (Mb)	Number of tags	highly correlated pairs		
			Number	Proportion	p-value
20	13-33	72	205	8.02%	0.0203
20	14-34	83	278	8.17%	0.0170
20	16-36	98	352	7.41%	0.0404
21	10-30	27	28	7.98%	0.0213
21	12-32	32	37	7.46%	0.0383
21	13-33	34	42	7.49%	0.0371
21	18-38	69	170	7.25%	0.0477
21	19-39	74	204	7.55%	0.0347
21	20-40	82	240	7.23%	0.0487
21	21-41	84	263	7.54%	0.0351
21	22-42	85	282	7.90%	0.0234
21	23-43	94	356	8.14%	0.0176
21	24-44	105	436	7.99%	0.0211
21	25-45	118	559	8.10%	0.0185
21	26-46	125	583	7.52%	0.0359

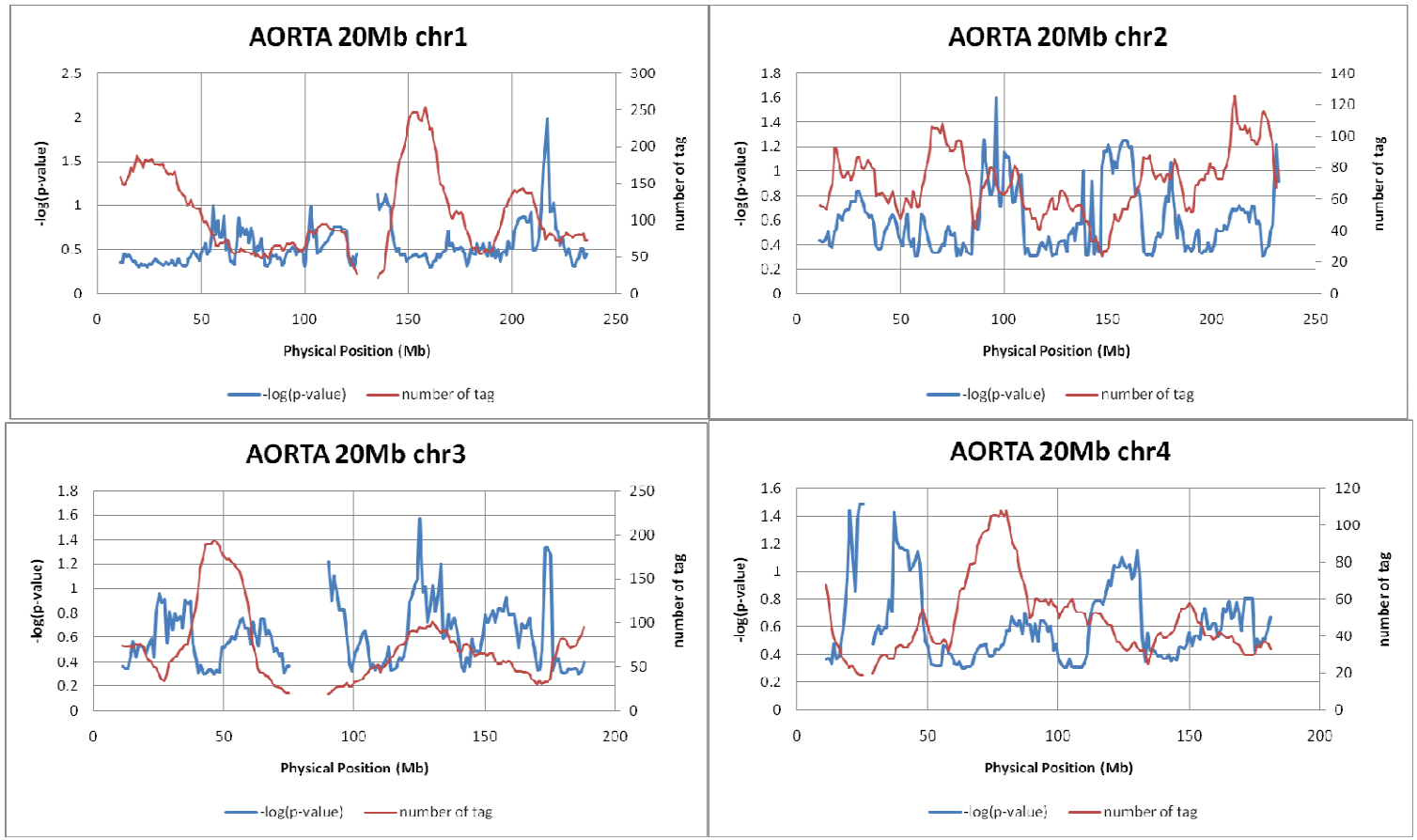


Figure 9: The results of high correlation analysis for the genome applied to the Aorta dataset. The blue curves show the distributions of logarithm of p-value of every region, and the red curves represent the number of tags in the region. The physical position of a region represents the midpoint of the range. For example, the analysis result from sampling window 115Mb-135Mb was plotted on physical position 125Mb.

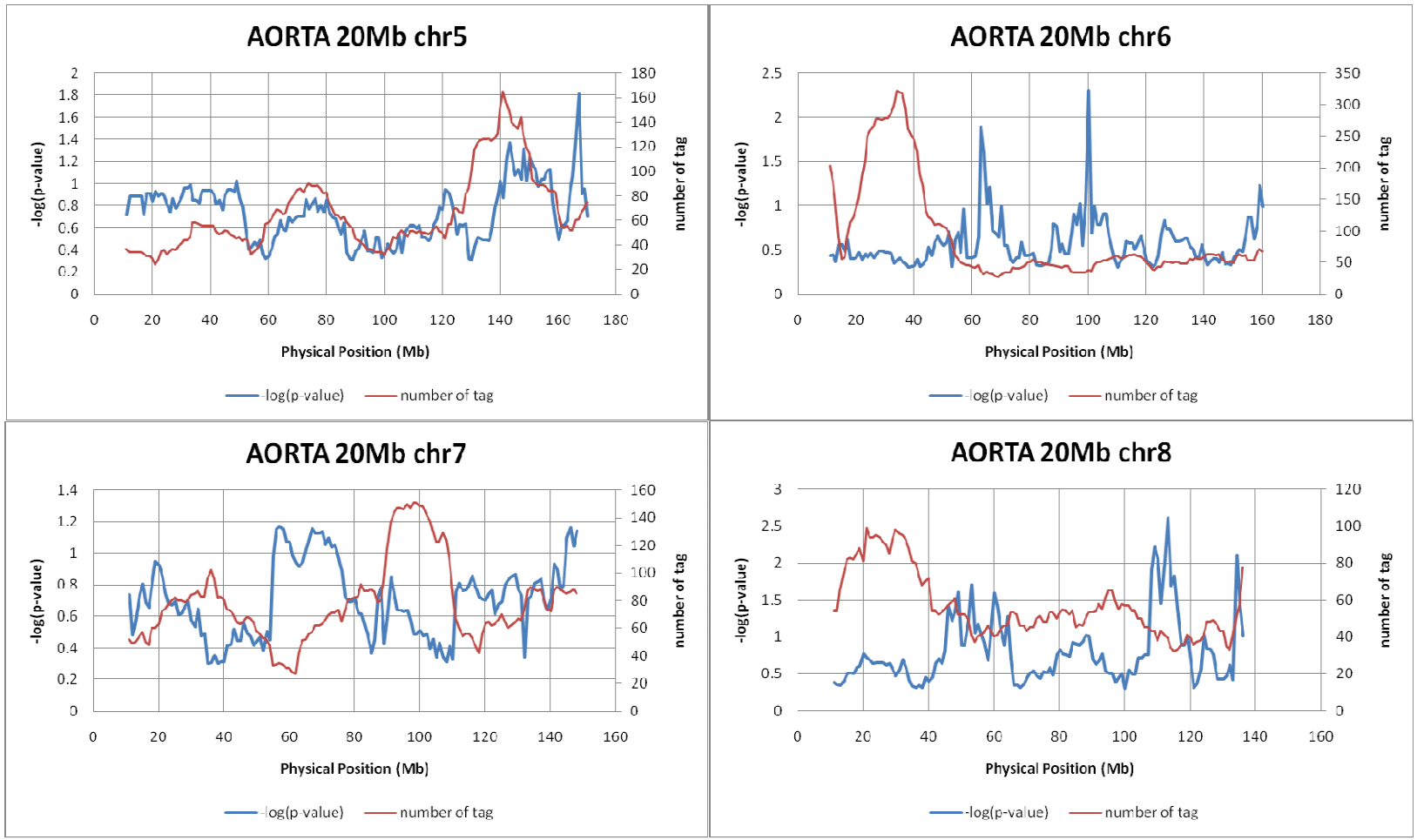


Figure 9: Cont'd.

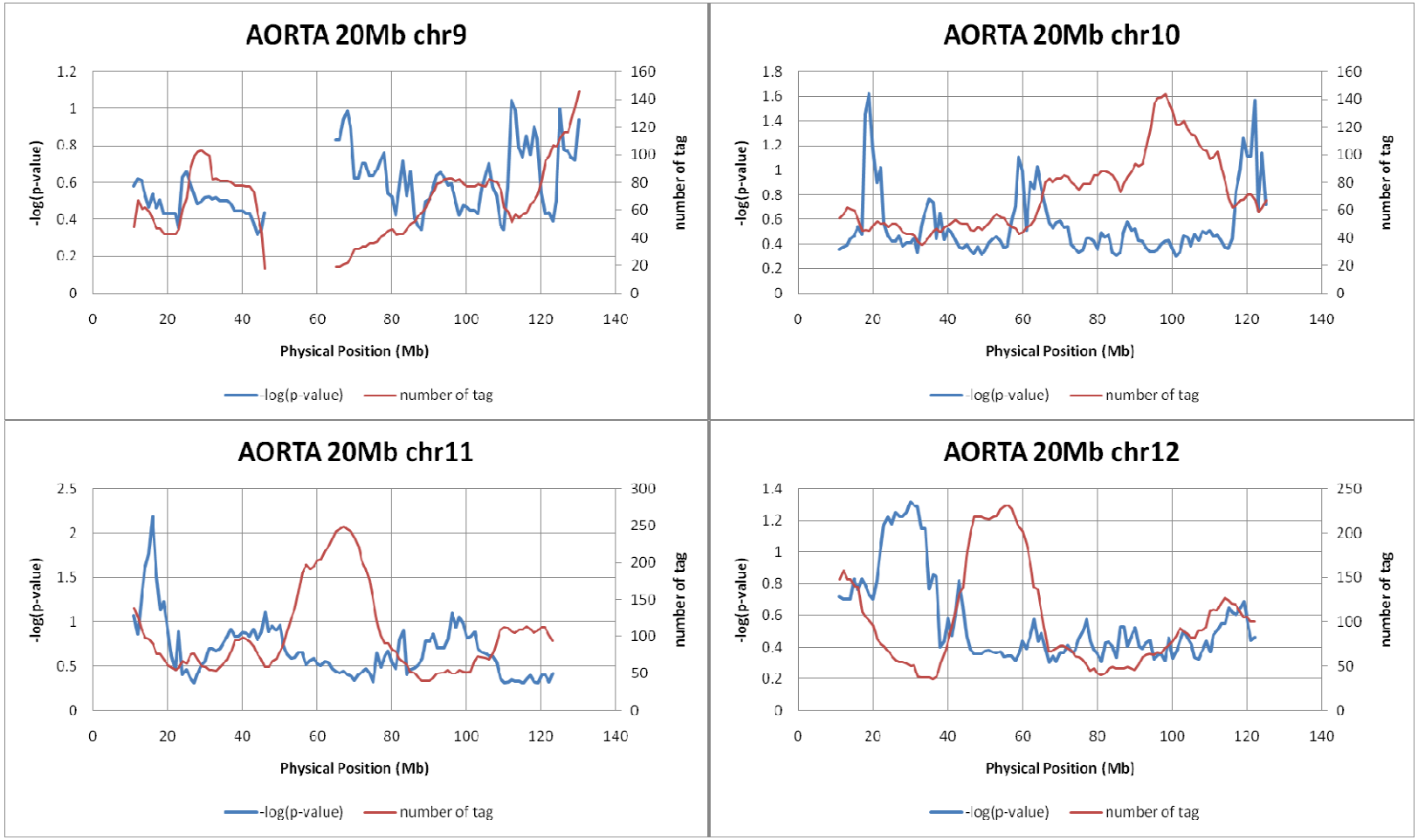


Figure 9: Cont'd.

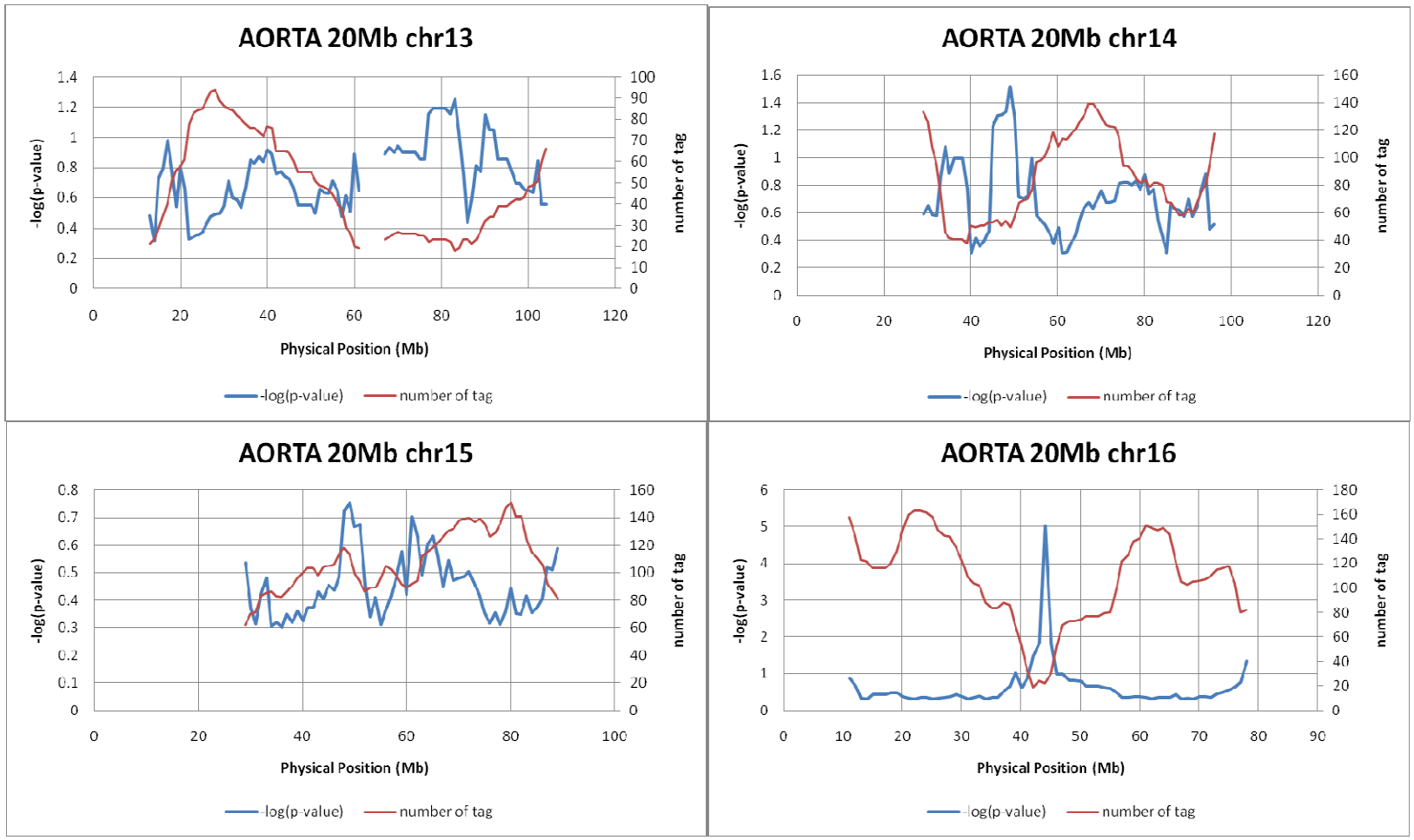


Figure 9: Cont'd.

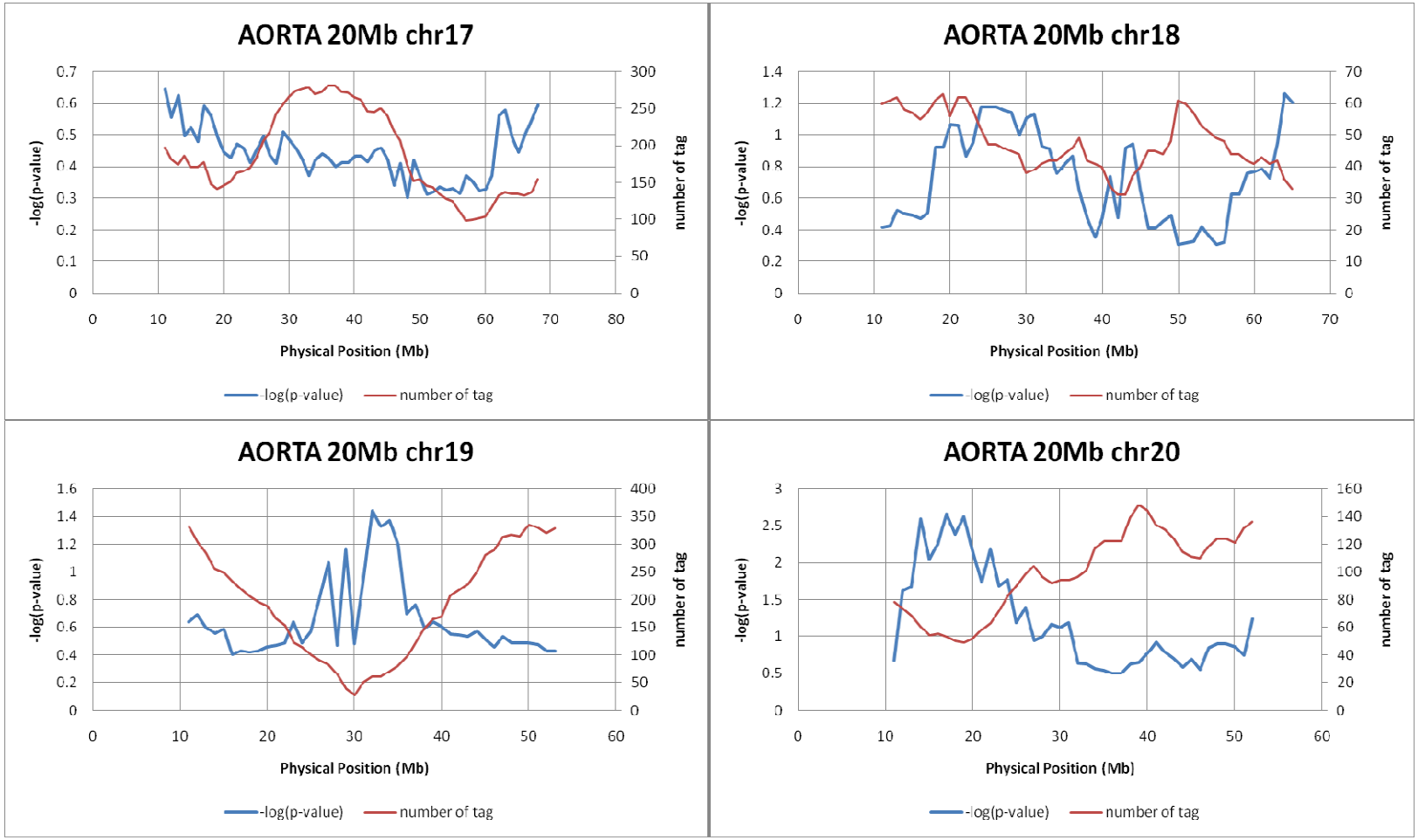


Figure 9: Cont'd.

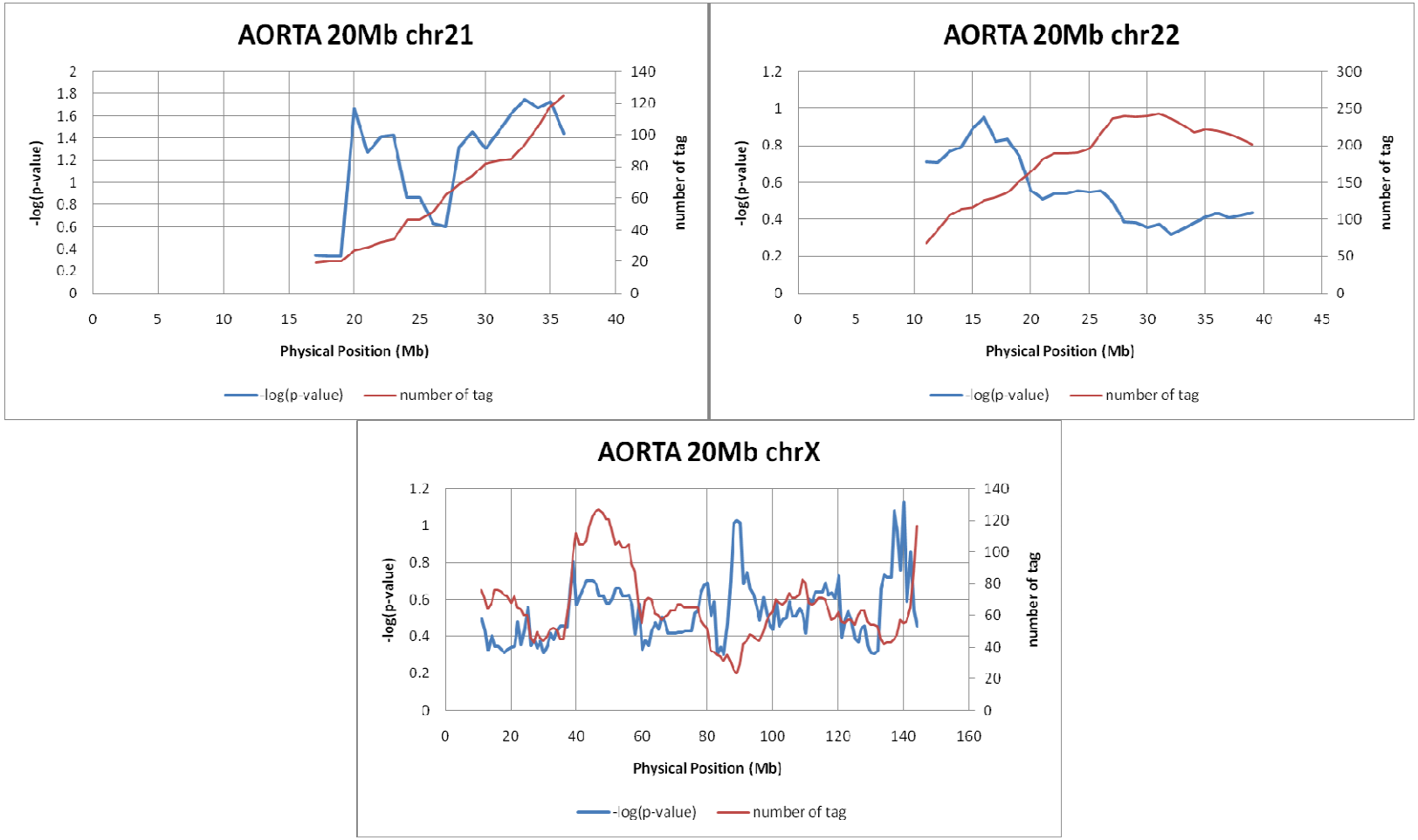


Figure 9: Cont'd.

3.6 Varying the threshold for the correlation coefficient

In order to detect pairs of microarray tags which showed high correlation, the threshold of absolute value of correlation coefficient ($\rho_{threshold}$) was set at 0.7 in previous analysis. Here we varied the threshold of correlation coefficient and repeated the correlation analysis across the genome in the Aorta dataset and the CATHGEN dataset. In our analysis, $\rho_{threshold}$ was set at 0.5, 0.6, 0.7, and 0.8, respectively. Figure 10 shows the statistical characteristics of *R* resulting from different threshold levels for the correlation coefficients.

The analysis results from the Aorta dataset are shown in Figure 11. For most of the chromosomes in the Aorta dataset, the distributions of proportion of highly correlated pairs tended to share coherent trends regardless of the threshold level of correlation coefficient. Shapes of these curves were quite similar, and the sites of peaked regions seemed independent of $\rho_{threshold}$. The statistical analysis of the CATHGEN dataset also showed the similarities between distributions from different threshold levels of correlation coefficients.

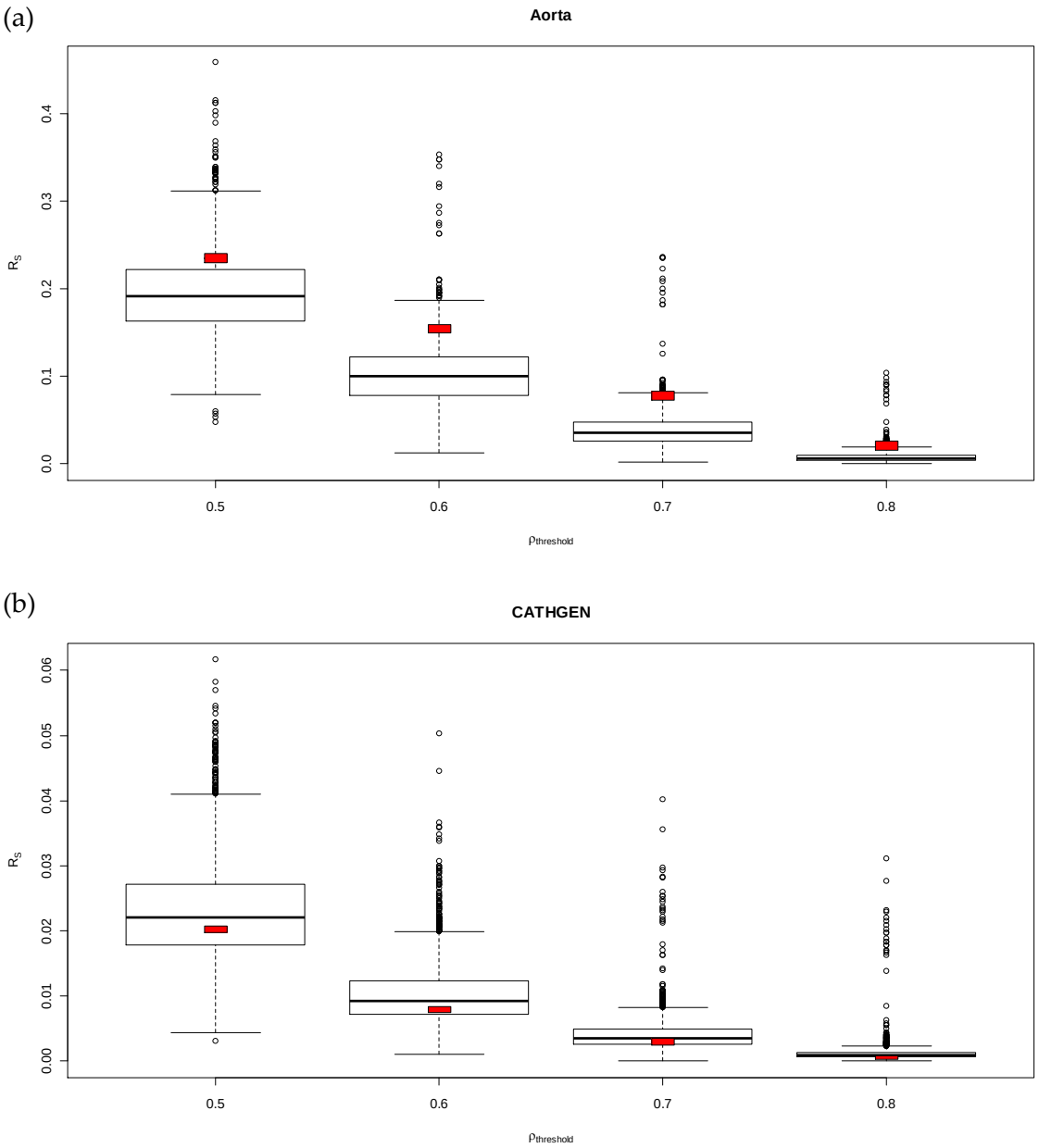


Figure 10: The statistical characteristics of the proportion of highly correlated pairs (*R*) resulted from various thresholds for the correlation coefficients from (a) the Aorta dataset and (b) the CATHGEN dataset. The *R* values from the regions with number of tags lower than 18 were eliminated. The red squares indicate the *R* values of 115Mb-135Mb on chromosome 3.

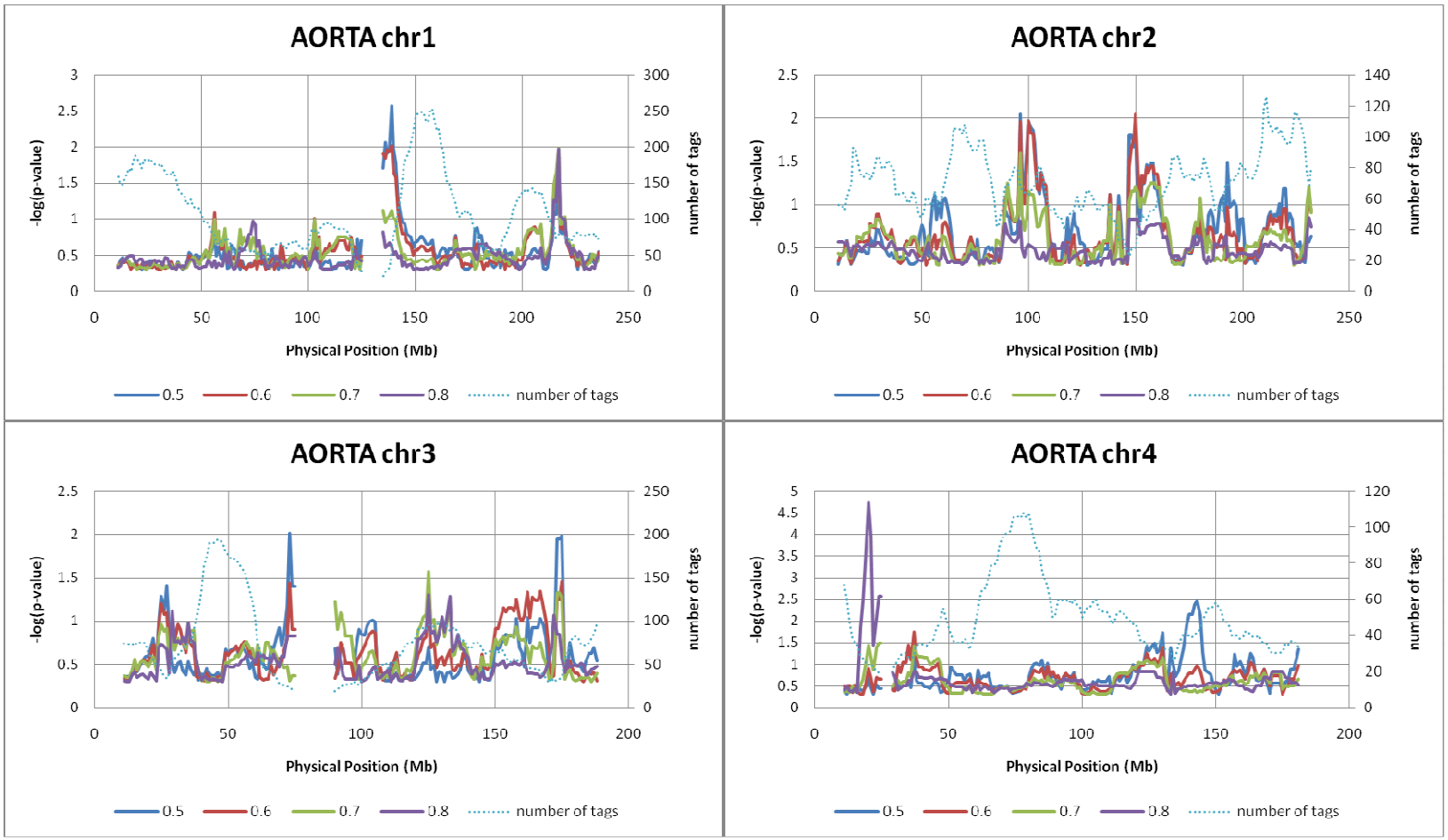


Figure 11: The distributions of proportion of highly correlated pairs of each chromosome from the Aorta dataset while varying threshold of correlation coefficient. The dotted curves represent the number of tags in the region for reference.

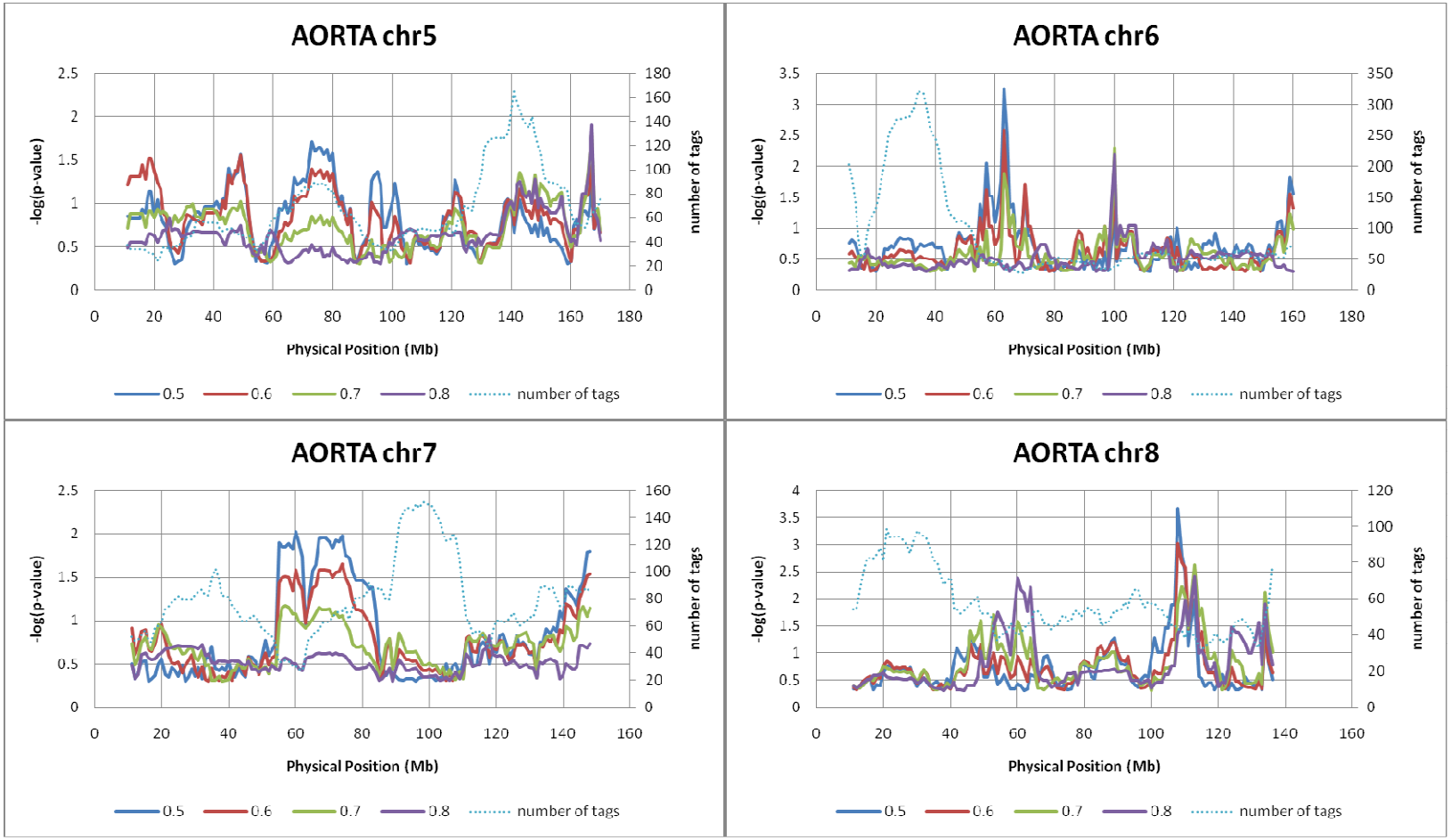


Figure 11: Cont'd.

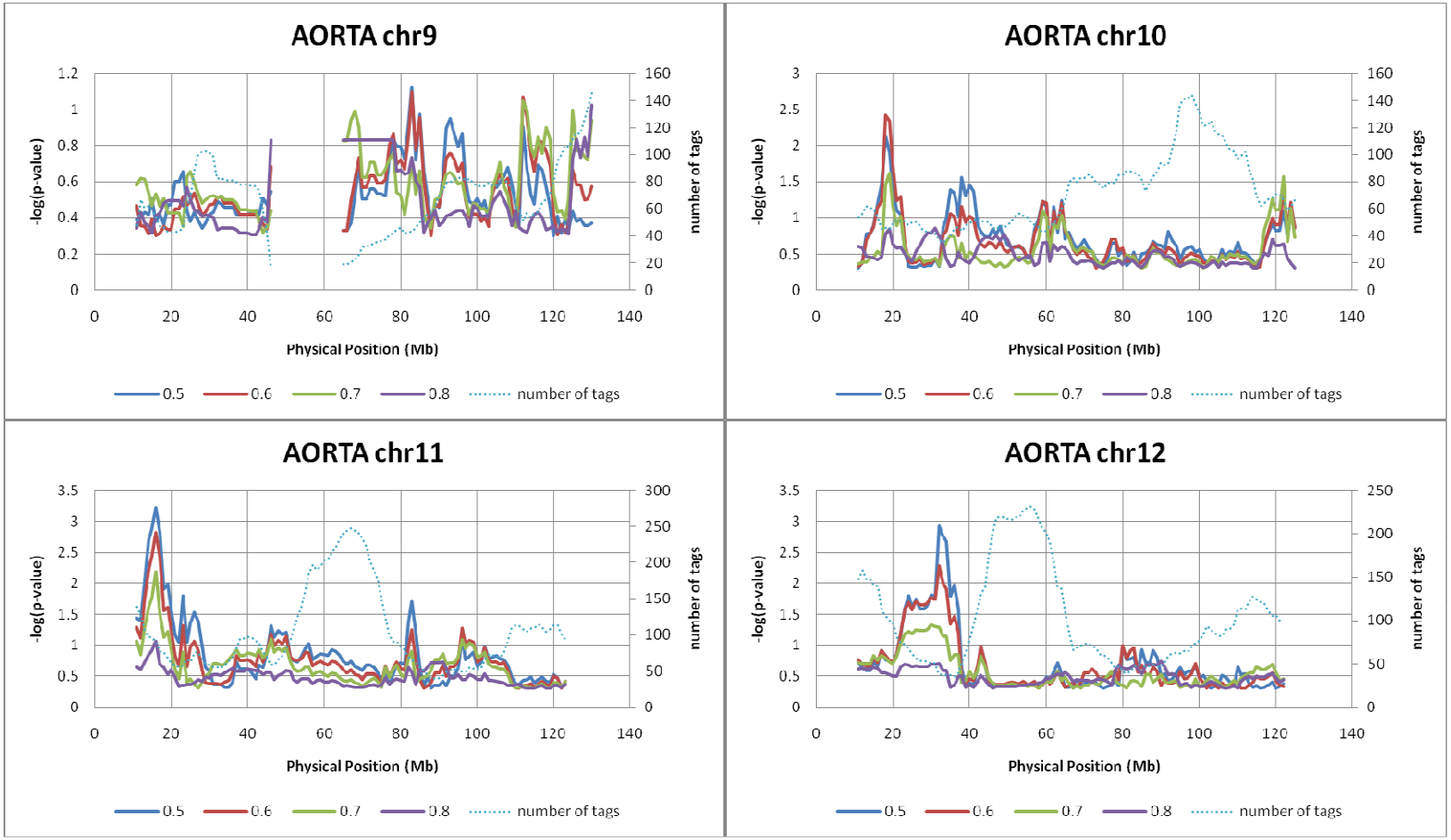


Figure 11: Cont'd.

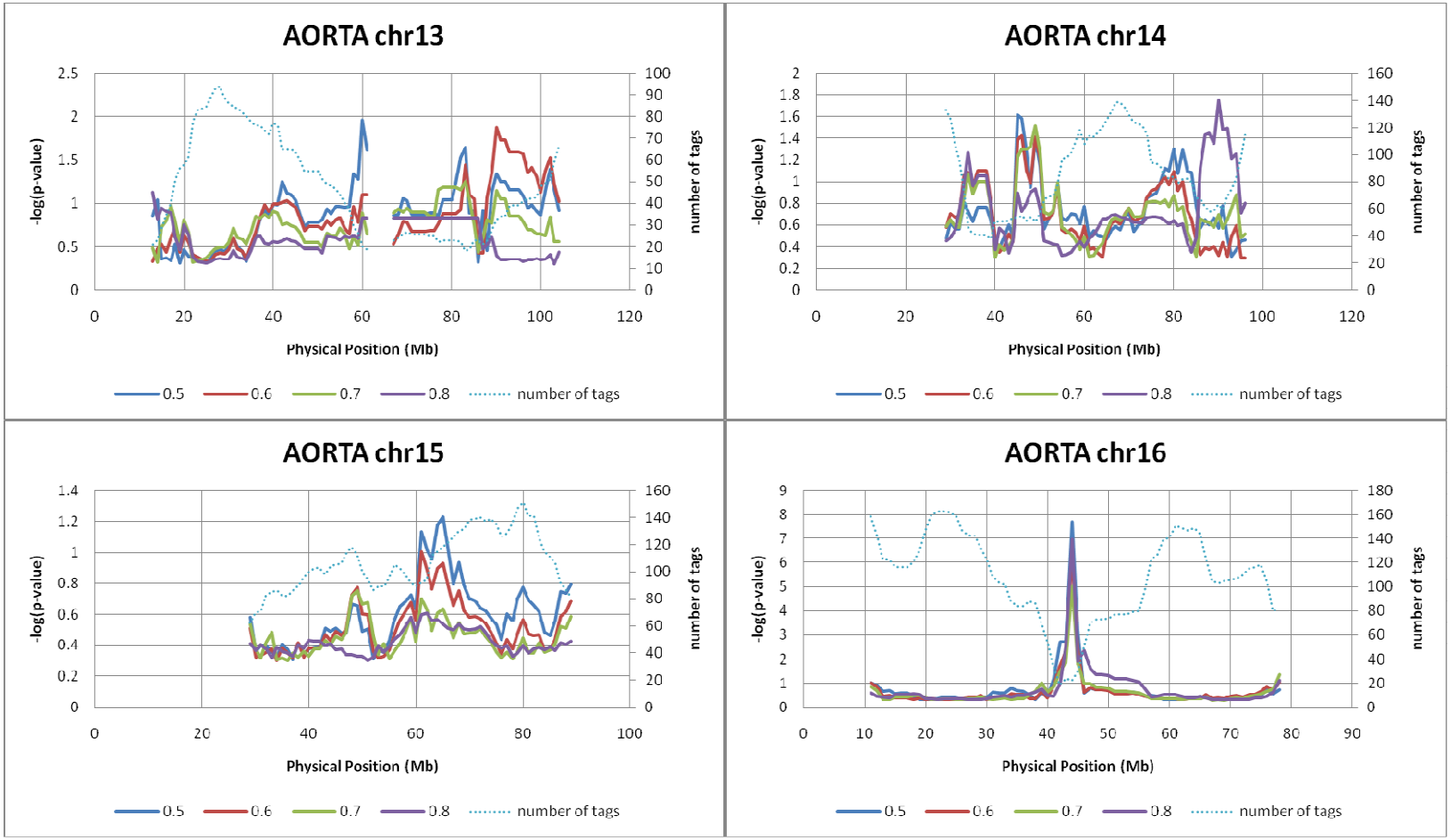


Figure 11: Cont'd.

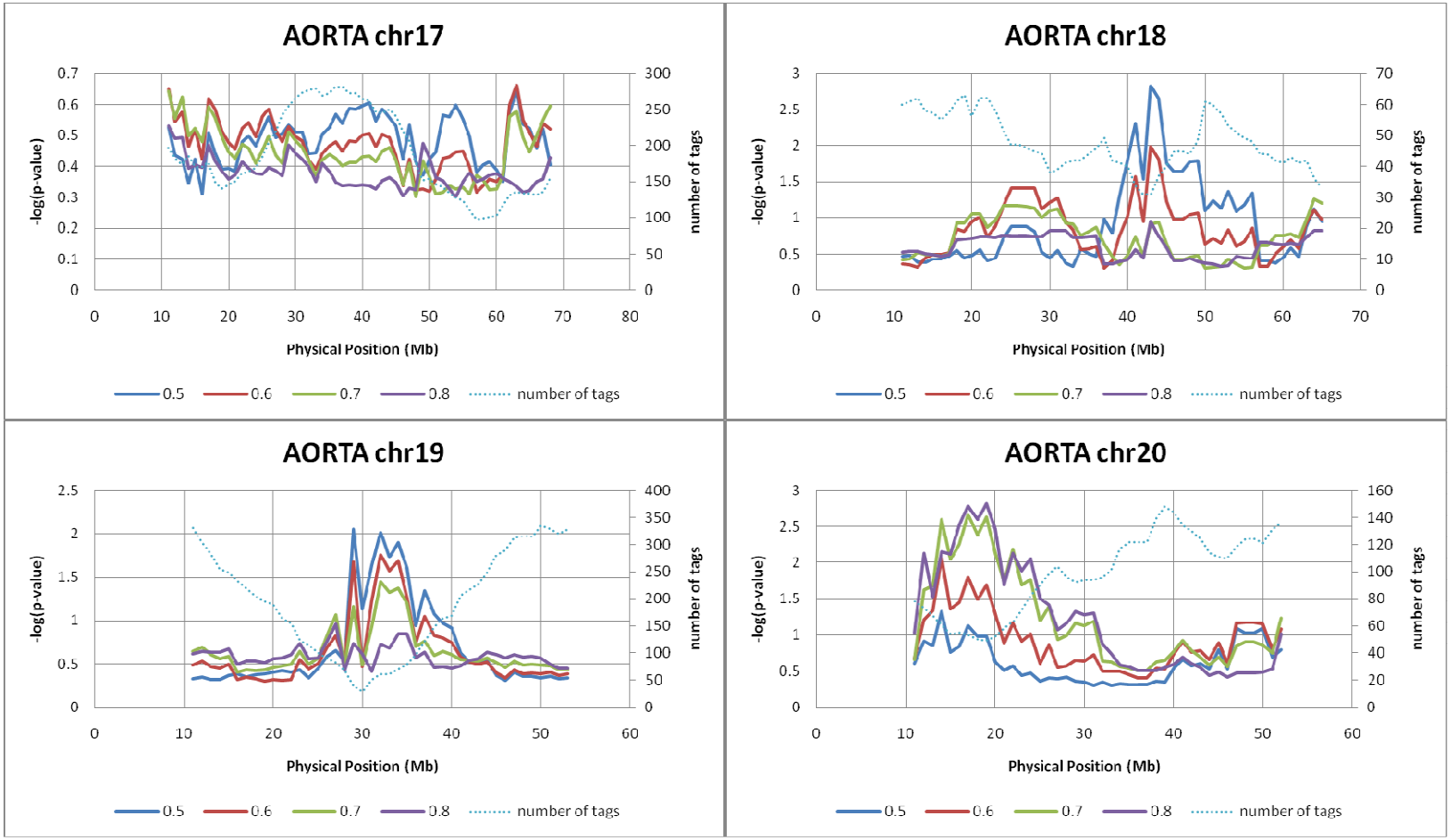


Figure 11: Cont'd.

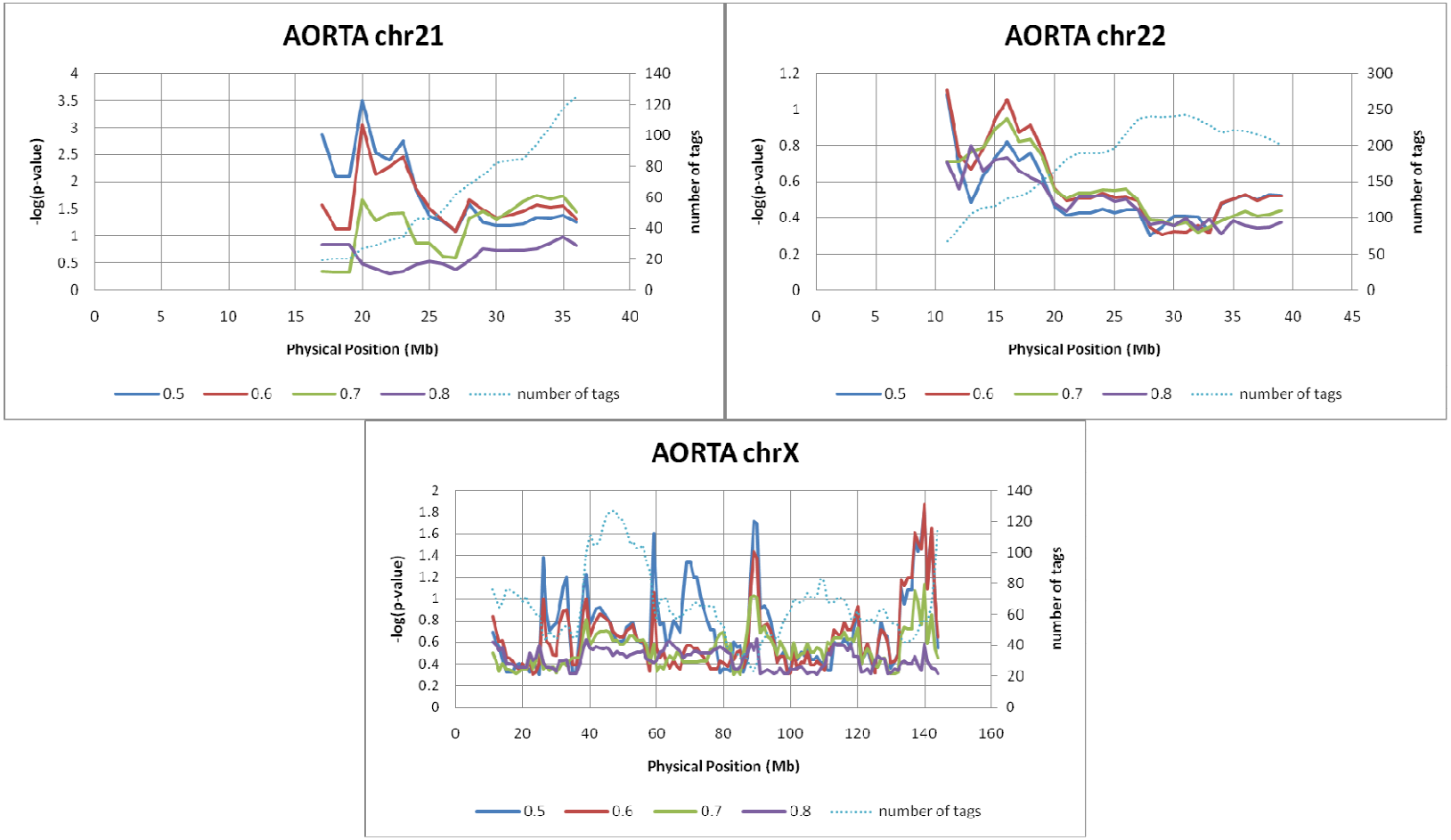


Figure 11: Cont'd.

3.7 Analyzing 115Mb – 135Mb on chromosome 3 from CATHGEN dataset

The highly correlated pairs from the CATHGEN dataset were obtained after applying correlation analysis as for the Aorta dataset. Figure 10 (b) shows the statistical characteristics of the proportion of highly correlated pairs (*R*) after analyzing the CATHGEN dataset. From the figure, the *R* value was lower than mean of *R* from the whole genome regardless of the threshold level. While applying $\rho_{threshold} = 0.7$, 338 pairs were obtained from the region 115Mb-135Mb on chromosome 3 (*R* =0.30%, *P* value = 0.3445 from Z-test). Figure 12 shows the comparison of distributions of proportion of highly correlated pairs on chromosome 3 between the Aorta dataset and the CATHGEN dataset. From Figure 12, there was no peak in the distribution at the region 115Mb-135Mb on chromosome 3 from the CATHGEN dataset.

3.8 Analyzing 115Mb – 135Mb on chromosome 3 from GDS1597 dataset

The highly correlated pairs from the GDS1597 dataset were also examined as the pairs from AORTA dataset. From the GDS1597 dataset, 595 pairs (*R* = 12.27%, *P* value = 0.1286 from Z-test) were obtained after applying $\rho_{threshold} = 0.7$. These 595 highly correlated pairs represented 71 genes in the region 115Mb-135Mb on chromosome 3. Furthermore, by comparison between lists of highly-correlated genes extracted from the Aorta dataset and GDS1597 dataset, several genes were overlapped between these

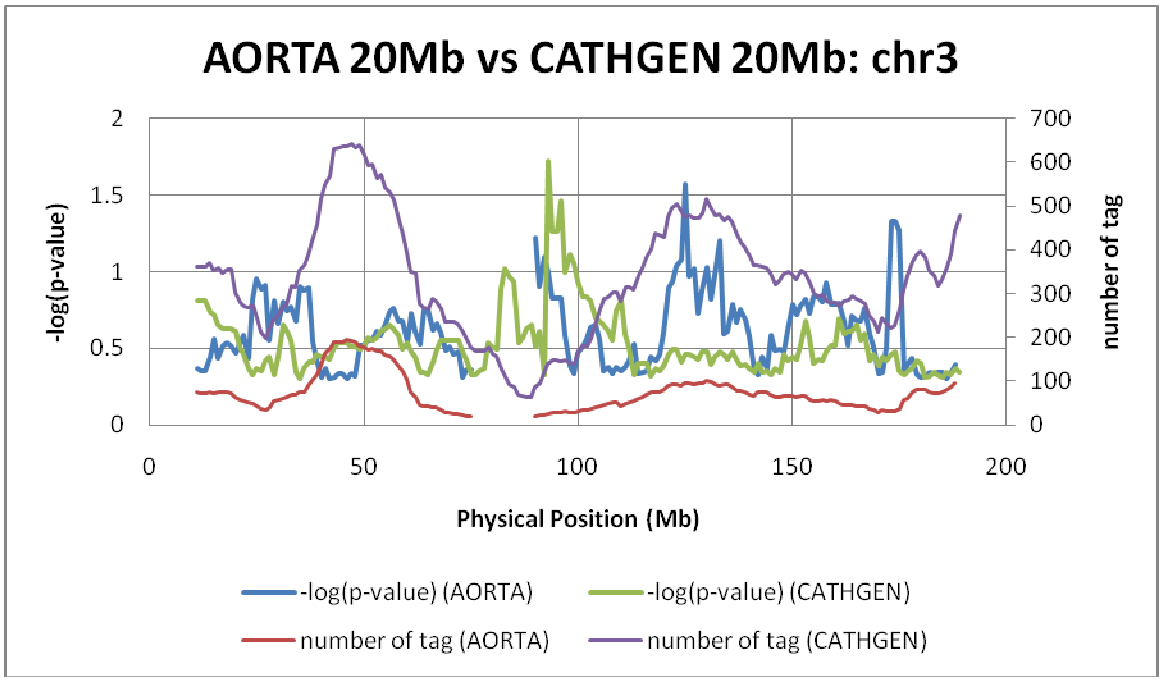


Figure 12: Comparison of distributions of proportion of highly correlated pairs on chromosome 3 between Aorta dataset and CATHGEN dataset. From the Aorta dataset, the distribution showed a peak in 115Mb-135Mb indicated by red arrow, but there was no peak at the same position from distribution come from the CATHGEN dataset.

two lists. Table 4 compares gene lists obtained only from the Aorta dataset, only from the GDS1597 dataset, and from both datasets. Twenty genes are observed from both the Aorta and the GDS1597 lists. Figure 13 demonstrates the comparison of distributions of proportion of highly correlated pairs on chromosome 3 between the Aorta dataset and the GDS1597 dataset. From Figure 13, there was also no obvious peak at chr3:115Mb-135Mb from the GDS1597 dataset.

Table 4: The list of genes appearing in highly correlated pairs. This table compares two lists of genes extracted from the Aorta dataset and the GDS1597 dataset. There are 20 genes appearing in both lists, 25 genes appear only in the gene list from Aorta dataset, and 51 genes appear only in the gene list from GDS1597 dataset.

Both (20 genes)	Aorta only (25 genes)	GDS1597 only (51 genes)	
ATP2C1	32119_at	ABTB1	NDUFB4
CASR	33953_at	ADPRH	NEK11
CDV3	38926_at	ASTE1	NPHP3
CNBP	ACPP	B4GALT4	NUDT16
DNAJC13	ATP6V1A	C3orf1	OSBPL11
DRD3	BFSP2	C3orf22	PARP15
GSK3B	BZW1	CCDC14	PIK3R4
GTF2E1	C3orf27	CCRL1	PLA1A
LSAMP	C7orf28A	CHCHD6	PLXNA1
MBD4	GATA2	COL29A1	PLXND1
PTPLB	GOLGB1	COX17	POPDC2
RPN1	GP9	CPNE4	QTRTD1
SLC15A2	H1FX	DNAJB8	RAB43
SNX4	HEG1	DTX3L	RABL3
TMCC1	HGD	EEFSEC	SEC22A
TOPBP1	ITGB5	FAM162A	SEMA5B
TRH	KIAA1539	GPR156	SLC12A8
ZBTB20	MRPL3	GPR175	TMEM108
ZNF148	MYLK	GRAMD1C	TMEM39A
ZNF80	NR1I2	IGSF11	UBA5
	PTMA	ISY1	UMPS
	RHO	KALRN	UROC1
	SEC61A1	KIAA1407	WDR5B
	STXBP5L	KPNA1	ZDHHC23
	UPK1B	MCM2	ZXDC
		MUC13	

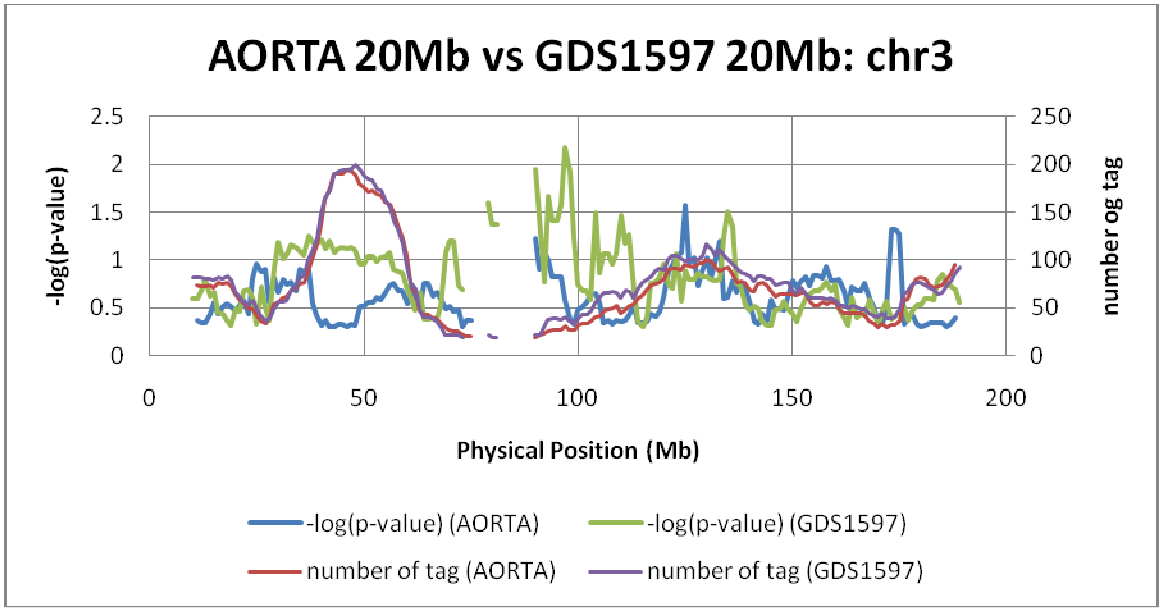


Figure 13: Comparison of distributions of proportion of highly correlated pairs on chromosome 3 between the Aorta dataset and the GDS1597 dataset. There was no peak at the same position in the GDS1597 dataset.

3.9 115Mb-135Mb on chromosome 3 showed more connections to the rest of the genome than others

Another interesting topic is whether the region 115Mb-135Mb on chromosome 3 also shows high connectivity to the rest of the genome. Here we applied a permutation test to verify that the region chr3:115Mb-135Mb (S) has more connections to the rest of the genome than other regions. After performing each permutation test on the Aorta dataset, the distribution of $\{\rho_{S_n S_{n'}}\}$ was significantly different from the distribution of $\{\rho_{SS'}\}$. The P value from Mann-Whitney U test comparing $\{\rho_{SS'}\}$ and $\{\rho_{S_n S_{n'}}\}$ was lower than 10^{-15}. Furthermore, the region S showed stronger connections toward S' than the

connectivity between S_n and S_n' which are generated by random sampling. Figure 14 (a) demonstrates that with larger $\rho_{threshold}$, the percentile of R_S tended to become higher in the distribution of R_{Sn} constructed by permutation test. Furthermore, Figure 14 (c) displays the distribution at $\rho_{threshold} = 0.7$ from Figure 14 (a). From Figure 14 (c), R_S was higher than most of R_{Sn} with $\mathrm{FDR} = 0.47\%$.

However, the phenomenon did not reappear in analysis of the CATHGEN data set. In Figure 14 (b), R_S was much lower than R_{Sn} from the permutation analysis of CATHGEN dataset.

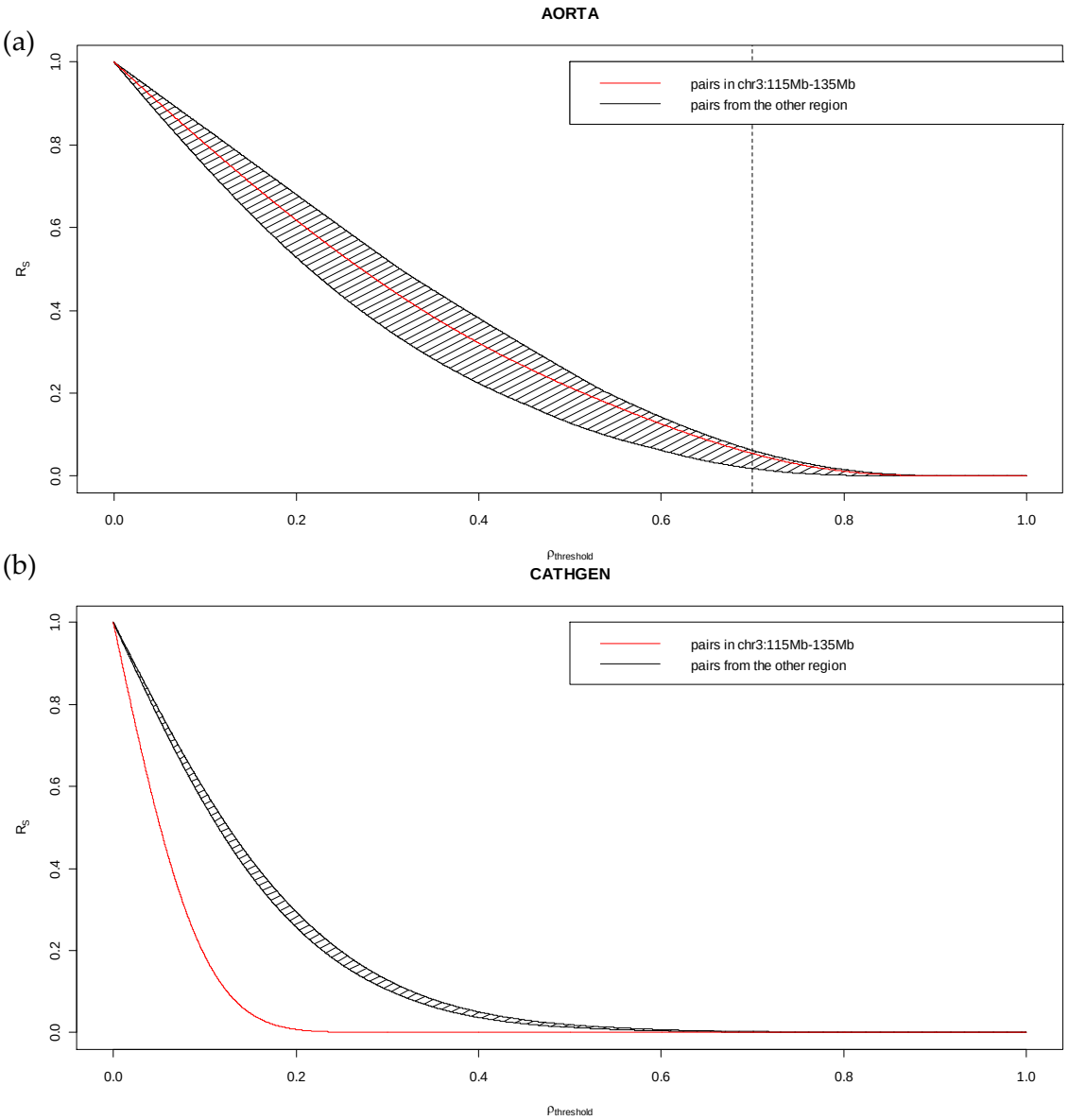


Figure 14 Complementary cumulative distribution functions of R_S. The figure shows variation of R_S by changing $\rho_{threshold}$. The black shadow demonstrates the distribution of R_{S1} to R_{SN}. The red line demonstrates proportion of highly correlated pairs between tags in 115Mb-135Mb on chromosome 3, *S*, and the others in the genome, *S'*. (a) The distribution displays the results from the analysis on the Aorta data set, and (b) shows the results from CATHGEN data set. (c) Distribution of proportion of high correlated pairs from randomly sampled tags. This figure comes from dissection of (a) at |ρ|=0.7. The false discovery rate is 0.47%. While analyzing the Aorta dataset, permutation number (N) is set at 3,000; while analyzing the CATHGEN dataset, N is set at 100.

(c)

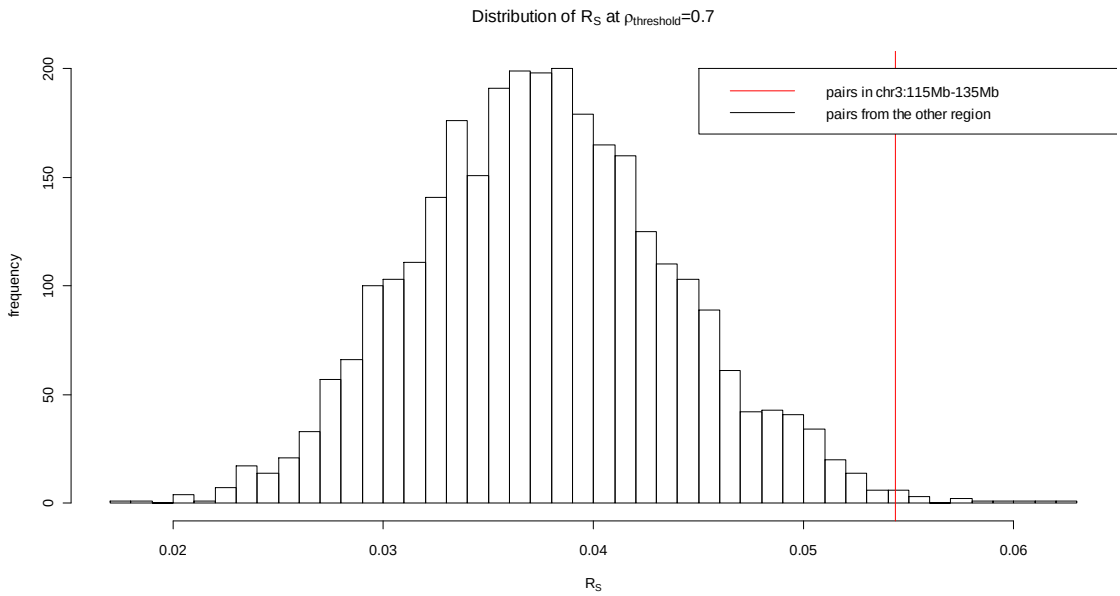


Figure 14: Cont'd.

4. Discussion

We have described a novel and unbiased genome-wide approach for examining for disease-associated regions in gene expression analysis of human aorta samples. The analysis is driven by correlation of expression of the clustered gene, rather than identifying genes whose expression level differs individually by an arbitrary amount. Our analysis not only identified the regions which we previously indicated relevant to atherosclerosis, but also unmasked several regions which might have association with the cardiovascular disease. Some of the genes in identified regions have significantly different genetic association odds ratio between patients and control group, and some of the genes are described in previous studies.

4.1 115Mb-135Mb on Chromosome 3

Our analysis suggested that the region 115Mb-135Mb on chromosome 3 (*S*) is particularly special in human aorta samples. After statistical analysis of all regions in the genome from the Aorta dataset, the region *S* is significantly different in our localized Pearson's correlation test with *P* value was 0.0266 from Z-test. By analyzing the Aorta dataset, 348 pairs with relatively strong correlation are constructed by 45 genes in the region *S*.

In order to verify the analysis accuracy of the genes associated with highly correlated pairs in the Aorta dataset, the same analysis procedures were applied on the GDS1597 dataset. The proportion of highly correlated pairs (*R*) in 115Mb-135Mb on

chromosome 3 was high (12.27%) although the *P* value was 0.1286 from Z-test. The relatively high *R* suggested that the region 115Mb-135Mb on chromosome 3 in the GDS1597 dataset also displays tight connectivity. Besides the considerable proportion of highly correlated pairs, many genes identified in the GDS1597 dataset were also involved in highly correlated pairs from the Aorta dataset. Table 4 lists that 20 genes appeared in both lists of highly correlated genes. The reason for the large proportion of overlapping genes might be tissue-specific factors of gene expression. The Aorta dataset and the GDS1597 dataset were sampled from aortas and coronary arteries in humans. The same sampling origins infer the similar location-oriented or organ-oriented physiological factors which affect gene expression. The gene expression affected by similar physiological environments might explain the similarity of connectivity and correlated genes.

In another point of view, we can verify the highly correlated genes by introducing genetic association analysis from patients with cardiovascular disease and controls. In the previous study, the DNA of some subjects in the Aorta dataset or in the CATHGEN project also have been genotyped. The odds ratio between cardiovascular disease patients and controls was computed and stored in the database of the Center for Human Genetics (CHG) [51]. By comparison between genes and SNPs, the possibility of a gene which was found to be highly correlated could be enhanced by having a significantly different allele frequency between cardiovascular disease patients and

Table 5: SNP-verification of genes involved in highly correlated pairs in the Aorta dataset or GDS1597 dataset. A gene was verified if there is at least one SNP with a significantly different (*P* value < 0.1) in proportion of minor allele in either the Aorta database or CATHGEN database.

Gene Origin	Verified	No Significant differences	No data	Total
Aorta + GDS1597	11	5	4	20
Aorta only	9	4	12	25
GDS1597 only	10	0	41	51
Total	30	9	57	96

controls. Table 5 lists the number of SNPs which have significantly different odd ratios between patients and controls from CHG SNPs database in the region 115Mb-135Mb on chromosome 3. Here we separated highly correlated genes into three groups: (1) genes involved in highly correlated pairs in both datasets, (2) only involved in the Aorta dataset, and (3) only involved in GDS1597 dataset. From table 5, eleven out of 20 (55%) genes involved in highly correlated pairs in the both datasets were observed having significantly different odd ratios between cardiovascular disease patients and controls (*P* value < 0.1).

In these eleven genes, LSAMP is the gene having most SNPs with significantly different odd ratios. LSAMP was first recognized as a cell membrane protein that guides the development of specific patterns of neuronal connections by acting as a selective homophilic adhesion molecule [52-54]. Recently, LSAMP has been discovered having tumor suppressor function with epithelial ovarian cancer by inhibits the proliferation of

tumor cells [55, 56]. However, LSAMP also has been found as a risk factor for developing atherosclerosis through regulating arterial wall smooth muscle cells [31].

In the list of genes involved in highly correlated pairs in both of the Aorta dataset and GDS1597 dataset, ATP2C1 and CASR are the genes related to calcium regulation in human cells. ATP2C1 would produce an ATPase-calcium transporter which was found related to regulate calcium concentration in Golgi apparatus. If ATP2C1 expression is altered, losing of Golgi Ca2+ pump could cause malfunction of Golgi apparatus and apoptosis [57, 58]. GOLGB1, a protein located on Golgi apparatus [59], also appeared as a gene involved in highly correlated pairs in the Aorta dataset. CASR produces a G protein-coupled calcium sensing receptor. The calcium sensing receptor can detect extracellular calcium concentration [60-63]. CASR is also involved in calcium concentration regulation in cells, related to processes as diverse as hormone secretion, gene expression, ion channel activity, modulation of inflammation, proliferation, differentiation, and apoptosis [64-66]. While undergoing atherogenesis, remodeling of arteries may involve local variations which change the physiological environments such as inflammation or calcium concentration. These factors could trigger these calcium-related pathways and raise the possibility of affecting cardiovascular disease. In addition, intracellular calcification within smooth vascular muscle and atheromata could change calcium homeostasis locally. With variation of calcium concentration, the risk of

having cardiovascular disease may raise with cooperation of ATP2C1 and CASR expression.

Among the SNP-verified genes, MBD4 is the most frequent gene appearing in the highly correlated pairs from both datasets. MBD4 is one of the subunits comprising nuclear protein related by presence in each of a methyl-CpG binding domain (MBD) [67, 68]. MBD4, acting as one of the human DNA repair proteins, recently has been indicated associated with cancer development [69-71]. Recently, several studies revealed the association between atherosclerosis and oxidative DNA damage [72-77]. MBD4 has been shown that it is able to repair thymine glycol which is a common DNA lesion produced by oxidative stress [78]. Disturbance of MBD4 balances could reduce the efficiency of DNA-repairing and increase the risk of atherosclerosis.

Finally, GTF2E1, ZNF148, ZBTB20, ZNF80 are transcription factors. These zinc finger proteins bind to DNA directly and change the transcription level to alter gene expressions. GSK3B has been shown as a risk factor of cardiovascular disease [79, 80]. GSK3B-deficient could also cause cardiomyoblast hyperproliferation [81]. Our data show that expressions of these genes are highly correlated, which could support that these genes are relevant to cardiovascular diseases.

Although some genes were only selected by high correlation from either the Aorta dataset or GDS1597 dataset, these genes still have been demonstrated that they were also related to cardiovascular disease. GATA2, an essential factor for the

development and differentiation of hematopoietic stem and progenitor cells, is a transcription factor that is indispensable for all kinds of hematopoiesis pathways. GATA2 is expressed in both endothelial cells and smooth muscle cells [29, 82-84]. Besides, KALRN, RHO, and MYLK are closely tied to each other within the Rho GTPase signal transduction pathway which regulates a wide range of cellular activity such as cell proliferation, cell migration, cell adhesion, cytoskeletal dynamics, oncogenic transformation, and inflammatory responses [2, 85-88]. Some of the Rho GTPases have been indicated in many cardiovascular activities, such as RhoA and its downstream effector Rho-kinase (ROCK) [2, 85, 89-92]. Our analysis also can retain these genes from genome by showing they are involved in highly correlated pairs in a 20Mb window.

Furthermore, we showed that LSAMP and GATA2 were highly correlated to each other with correlation coefficient was 0.7846 in the Aorta dataset. From Figure 4, we can observe that gene expressions of LSAMP and GATA2 are highly positive correlated. GATA2 has been association with cardiovascular disease in early-onset patients [29], and LSAMP has been demonstrated to be associated with later-onset cardiovascular disease [31]. The genomic distance between LSAMP and GATA2 is over 12Mb. The large distance between these two genes implies that they are not likely to have a shared promoter. The high relevance suggested that there are some mechanisms which are still unknown linking these two genes. In addition, from Figure 5 we can discover that the gene expressions related to GATA2 and LSAMP were mostly *cis*-directed. The *cis*-

directed relationships between GATA2 and LSAMP represent that if gene A is highly correlated to GATA2 with a positive correlation coefficient, then gene A would be also correlated to LSAMP with a positive correlation coefficient. Figure 15 demonstrated the connections aligned with GATA2 and LSAMP. These connections inferred that there should be a gene network underlying these genes. The gene network would integrate several different pathways and regulate these risk factors coherently. Our analysis linked GATA2 and LSAMP together and suggested that these two genes might be functionally related.

Currently we have demonstrated that the genes are highly connected in the region 115Mb-135Mb on chromosome 3 (*S*). Furthermore, our permutation tests showed that the genes in the region *S* were not only highly connected within the region, but also the genes were highly connected to the other regions than other sets of randomly selected genes. From figure 14(c), we could observe that with a correlation coefficient threshold of 0.7, proportion of highly correlated pairs ($R_{SS'}$), by which the pairs are constructed by the every possible combination between the region S and its complementary region S′, is higher than randomly selected gene sets with false discovery rate is 0.47%. The results of the localized Pearson's correlation tests suggested the high connectivity within the region *S*, and the results of permutation tests inferred that connectivity between the region *S* and its complement *S′* is also significantly higher

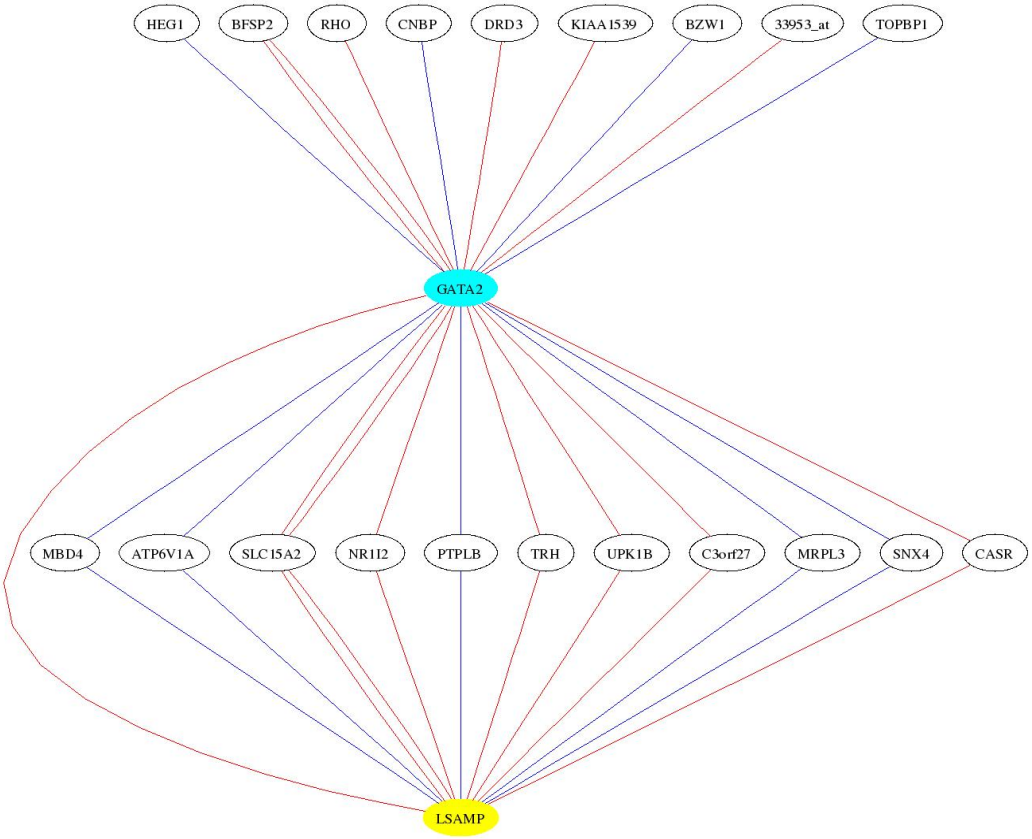


Figure 15: Connections between genes involved in highly correlated pairs more than 10 times. The red edges represent positive correlation coefficients, and the blue edges represent negative correlation coefficients.

than others. This phenomenon hints that the region 115Mb-135Mb on chromosome 3 might be an activation center related to cardiovascular disease. The genes in the region *S* would have linkage to genes in other regions and regulate cardiovascular disease-associated phenotypes.

The importance of the region 115Mb-135Mb on chromosome 3 for cardiovascular disease was also been supported by copy-number variation analysis. Figure 16 demonstrated the reported copy-number variation from Database of Genomic Variation

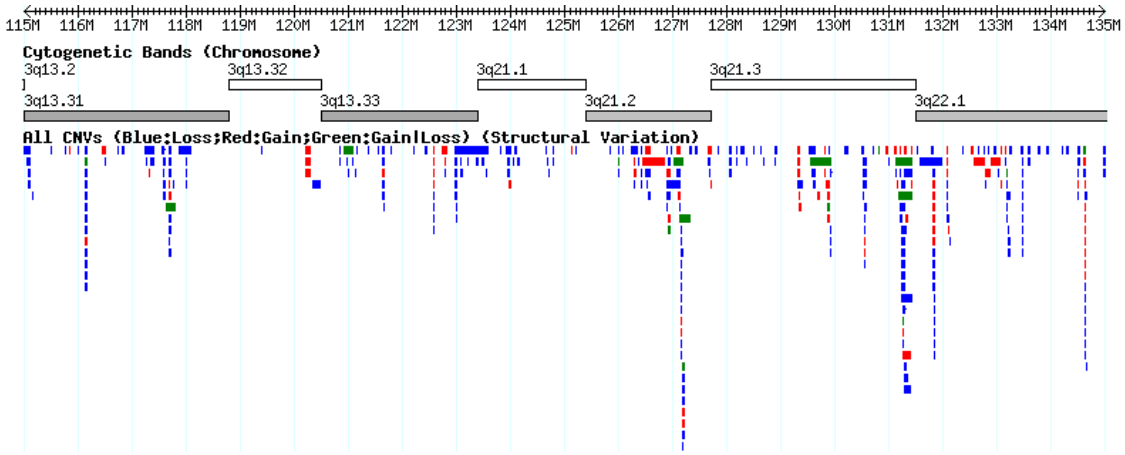


Figure 16: Copy-number variations reported in 115Mb-135Mb on chromosome 3. The red blocks indicate reported gain segments, blue blocks indicate loss segments, and green blocks indicate variations (could be gain or loss).

[93]. From the database, we discovered that ATP6V1A, ZDHHC23, ZBTB20, LSAMP, NR1I2, GSK3B, STXBP5L, SLC15A2, CASR, KALRN, ITGB5, SLC12A8 , HEG1, SNX4, ZNF148, UROC1, ZXDC, CHCHD6, MCM2, EEFSEC, GATA2, DNAJB8, C3orf27, RPN1, PLXND1, TMCC1, PIK3R4, ATP2C1, NEK11, NUDT16, CPNE4, TMEM108, BFSP2 and CDV3 were reported having deletion or duplication in copy number [13, 93-115]. The variations implied that the copy-number variation might be one reason why the expression levels of these genes were highly correlated because many of the copy-number variations occurred simultaneously.

In summary, the localized Pearson's correlation tests applied on the Aorta dataset and GDS1597 dataset have shown that the region 115Mb-135Mb on chromosome 3 is significantly connected within the region. In addition to the proportion of highly correlated pairs, the genes involved in highly correlated pairs have been indicated that

they were associated to cardiovascular disease by previous genetic linkage studies or SNP-association studies [116]. Furthermore, by tighter connectivity than other region, the region 115Mb-135Mb on chromosome 3 could regulate gene in other regions to express coherent phenotypes.

4.2 The Regions Detected by the Localized Pearson's Correlation Test

After the region 115Mb-135Mb on chromosome 3 was analyzed successfully by the localized Pearson's correlation test, we applied the analysis methodology on the whole genome. By dividing the entire genome into smaller windows, the possible combinations of microarray tags were reduced to a reasonable size to allow efficient computation of genome-wide measures. However, because of different density of microarray tags, the number of tags in a given window could be too small to provide a convincing analysis. For example, in the windows near the centermere of chromosome 3, the numbers of tags in those windows are less than 18, which resulted in the numbers of all possible pairs less than 150. With a small denominator, any slight change in numerator – *i. e.* the number of highly correlated pairs – could cause dramatic alteration in proportion of highly correlated pairs (R). Figure 7 shows the cumulative density function of proportion of highly correlated pairs (R) from the Aorta dataset. It is easy to observe that the regions with low number of tags ($N < 18$) tended to have extreme R-value. Eighteen regions with R-value = 0 all belong to regions with small number of tags. The regions with R-value greater than 10% belong to regions with small number of tags,

chromosome Y, and chr16:34Mb-54Mb which has only 22 tags in the window. The *R*-value distribution suggests that the type I error rate might be high if we accept all regions regardless of the number of tags. After applying the threshold of tag counts on our analysis, the extreme *R*-value caused by tag-number effects were removed, and the retained regions demonstrated a normal distribution of *R*-value as in figure 8 (b). We also applied another way to account for the various number of microarray tags. By fixing the window size at 95-tag long, the number of tags was guaranteed to be long enough. From figure 8 (c), we could discover that the distribution of *R*-value was closer to normal distribution than figure 8 (b). However, tag-oriented dissection would cause the variation of window size. The range of tag-oriented windows is between 600kb to 62M with mean = 20.4Mb and standard deviation is 12.4Mb. Huge variation of length of window would disturb the "localized properties" of sampling window. For example, tags on chromosome Y all belong to one window because the tags on chromosome Y are less than 95. After comparing the pros and cons of region sampling strategies, we decided to fix the length of window at 20Mb and applied a threshold on number of tags in a region. However, in general, the results of our study were quite stable comparing the fixed region length with a threshold to the fixed tag length.

Our analysis in the Aorta dataset showed there are 22 integrated regions having significantly high connectivity with P value < 0.05. The region chr3:115Mb-135Mb is one of the regions. We also tried to use our SNPs database to verify these regions. However,

most of these regions did not have enough SNPs genotyped in the target genes to verify the genes involved in highly correlated pairs by genetic association. Table 6 shows the genes which were verified by DUMC SNPs database. From Table 6, the q-terminal of chromosome 5 had 9 genes involved in highly correlated pairs have been genotyped and showed significantly different odds ratio (*P* value < 0.1) between cardiovascular disease patients and controls. The q-terminal of chromosome 5 has been also revealed having

Table 6: Genes with significantly different odd ratios between cardiovascular disease patients and controls in DUMC SNPs database. These genes were also participated in highly correlated pairs from the Aorta dataset. Most genes involved in highly correlated pairs could not be verified by SNPs since many genes have not been genotyped in the case-control series.

Chr	Region (Mb)	Genes verified by SNPs odd ratio				
1	205-227	CR1	TP53BP2			
2	86-106	IGKC				
3	163-184	N/A				
4	13058	N/A				
5	133-158	CD74	CSF1R	CTNNA1	IK	
5	156-177	FABP6	FGF18	LCP2	NKX2-5	STK10
6	53-74	N/A				
6	90-110	N/A				
8	36-71	N/A				
8	98-126	N/A				
8	124-145	SLA				
10	8-29	N/A				
10	112-132	FGFR2				
11	4-27	RRAS2				
14	36-59	N/A				
16	32-55	N/A				
16	68-88	PSMD7				
19	22-44	N/A				
20	2-36	N/A				
21	10-46	CSTB				

significantly high LOD score in the region 5q31 with P value < 0.081 [28]. The region 3q13 was also supported by this study [28]. However, the other regions (1q25, 7p14, and 19p13) with linkage peaks in the study [28] were not selected from our genome-wide expression scan.

Now we focus on the genes verified by significant SNP odds ratio between cardiovascular disease patients and controls. These genes were genotyped because they were either in region of linkage (*e. g.* chromosome 5) or were differentially expressed [20]. Complement Receptor I (CR1) has been shown to be relevant to coronary artery disease by gene knock-out studies and genotype studies [117-123]. Increased CD74 expression would enhance forming atherosclerosis plaque by contributing to inflammatory responses in vascular cells [124]. FGF18 and FGFR2, the fibroblast growth factors, might be associated with the process of atherosclerosis. NKX2 is an essential transcription factor regulating development and formation of heart [125-128]. Cystatins recently was found as a risk factor of cardiovascular disease if it lost its function [129]. Although not many genes were verified by SNPs data in DUMC SNPs database, many verified genes were associated with cardiovascular disease supported by previous studies.

We examined the robustness of our results to the choice of correlation coefficient threshold. Applying different correlation thresholds did not qualitatively alter the analysis results. From Figure 11 shows that the positions of peaks were consistent

regardless of the choice of correlation coefficient threshold between 0.5 and 0.8. We could also observe some regions with a high *R*-value but not high enough to pass the threshold 0.7. For example, chr2q22-24 was one of these regions. Previous studies showed that loss of heterozygosity on hMSH2 (2p22), hPMS1 (2q24), and hMLH1 (3p21) would cause atherosclerosis [130]. Chr3p22.3 has been identified as a new coronary artery disease risk locus [116]. Chr4q28 has been identified as a risk factor of coronary and extracoronary atherosclerosis [131]. The results inferred that peak positions of *R*-value were consistent if we set the correlation coefficient threshold higher than 0.5 to guarantee highly correlated pairs would be retained.

4.3 Factor of different results might be Tissue-specific

By analyzing the Aorta dataset and the GDS1597 dataset, we identified that the regions with high proportions of highly correlated pairs were relevant to cardiovascular disease. However after analyzing other datasets, including the CATHGEN, the GSE2014, the GSE6088, and the GDS268 dataset, these regions did not show high *R*-values in the region 115Mb-135Mb on chromosome 3. In contrast, the *R*-values in these regions were even lower than 0.5%. This phenomenon might result from tissue-specific factors. The RNA samples from these dataset were not sampled from the aorta or artery. The CATHGEN dataset and the GSE6088 cells were sampled from peripheral blood, which could have quite different physiological environments from coronary artery or aorta. The GSE2014 dataset sampled from the left ventricle free wall and the GSE268 dataset

sampled from abdominal skeletal muscle cells could also have local factors which could influence local cells without affecting gene expressions in aorta or coronary arteries. Nevertheless, the result from the GSE13985 dataset was an exception. The GSE13985 dataset was sampled from peripheral blood, but the analysis results were close to the results from the Aorta dataset and GDS1597 dataset which were sampled from aorta. The proportion of highly correlated pairs was approximately 5% which is much higher than the other datasets sampled from peripheral blood. In addition, the list from the GSE13985 dataset of genes involved in highly correlated pairs was similar to the lists from the Aorta dataset and the GDS1597 dataset (Appendix B). The results inferred that there might be some universal factors which affect gene expression levels but in some case the tissue-specific factors might be dominant.

4.4 Conclusion

We conducted a novel analysis to put genomic localization information about gene expression along with genetic association. However, we are unable to examine technical factors or quality differences while hybridizing the samples. Some varieties of gene expressions might result from the differences between microarray platforms, including number of probes and distribution of tags. In addition, the lack of information of sample heterogeneity means that we cannot be certain that whether correlations are stronger in cardiovascular disease patients than controls or not. Notwithstanding its limitations, the highly correlated regions not only reveal the genes which may be

associated with cardiovascular disease, but also suggest that these regions might have some special reasons that are associated to the disease more than other regions. A greater understanding of the genetic basis of the disease could potentially lead to the development of novel preventive or therapeutic strategies for cardiovascular disease.

Appendix A

The table below lists detailed verification of genes involved in highly correlated pairs in the region of 115Mb to 135Mb on chromosome 3 in the Aorta dataset or GDS1597 dataset. If there is at least one SNPs having significantly different (P value < 0.1) in proportion of minor allele in either the Aorta database or CATHGEN database, the gene would have dark background and be counted as verified in Table 5. Total SNPs columns represent the number of SNPs located in the gene have been genotyped.

Gene symbol	Number of highly correlated pairs involved in		Number of SNPs					
			Aorta database			CATHGEN database		
	Aorta dataset	GDS1597 dataset	*P* value < 0.05	*P* value < 0.1	Total SNPs	*P* value < 0.05	*P* value < 0.1	Total SNPs
SLC15A2	43	6	0	0	15	0	0	15
CASR	41	9	1	1	5	0	2	6
MBD4	32	30	0	0	7	0	2	7
SNX4	31	34	0	0	0	0	0	1
TOPBP1	27	30	0	0	1	0	0	1
DRD3	22	13	0	0	0	0	0	1
CDV3	20	23	0	0	1	0	0	1
PTPLB	17	34	0	0	1	0	0	2
TRH	14	2	0	0	0	0	0	0
LSAMP	14	1	12	17	90	32	46	113
DNAJC13	13	1	0	0	1	0	0	1
CNBP	11	28	0	0	0	0	0	0
GTF2E1	7	32	0	1	1	0	0	1
RPN1	4	35	0	0	2	1	2	6
ZNF148	3	59	1	1	1	0	0	2
ZBTB20	2	56	1	1	6	1	4	12
ATP2C1	1	38	3	4	4	1	4	7
TMCC1	1	20	0	0	0	4	4	5
ZNF80	1	14	0	1	4	0	0	5

Gene symbol	Number of highly correlated pairs involved in		Number of SNPs					
			Aorta database			CATHGEN database		
	Aorta dataset	GDS1597 dataset	*P* value < 0.05	*P* value < 0.1	Total SNPs	*P* value < 0.05	*P* value < 0.1	Total SNPs
GSK3B	1	1	0	1	4	1	1	4
RHO	43	0	0	0	0	0	0	0
HEG1	40	0	0	0	0	0	0	2
BFSP2	39	0	1	1	6	0	1	7
KIAA1539	31	0	0	0	0	0	0	0
C3orf27	30	0	0	1	1	0	0	2
UPK1B	27	0	0	0	1	0	0	1
NR1I2	24	0	0	0	1	0	0	1
GATA2	23	0	0	1	21	3	9	23
ATP6V1A	20	0	0	0	0	0	0	1
33953_at	19	0	0	0	0	0	0	0
GP9	17	0	0	0	0	0	0	0
MRPL3	17	0	0	0	1	0	0	2
BZW1	14	0	0	0	0	0	0	0
HGD	12	0	0	0	1	0	1	1
SEC61A1	7	0	0	0	1	1	1	1
STXBP5L	5	0	0	0	3	1	1	7
38926_at	4	0	0	0	0	0	0	0
ITGB5	4	0	0	0	2	0	0	6

Gene symbol	Number of highly correlated pairs involved in		Number of SNPs					
			Aorta database			CATHGEN database		
	Aorta dataset	GDS1597 dataset	*P* value < 0.05	*P* value < 0.1	Total SNPs	*P* value < 0.05	*P* value < 0.1	Total SNPs
32119_at	2	0	0	0	0	0	0	0
ACPP	1	0	0	1	1	0	0	2
C7orf28A	1	0	0	0	0	0	0	0
GOLGB1	1	0	0	0	16	0	1	16
H1FX	1	0	0	0	0	0	0	1
MYLK	1	0	7	9	40	17	28	52
PTMA	1	0	0	0	0	0	0	0
CCDC14	0	36	0	0	0	0	0	0
PLXNA1	0	35	0	0	0	0	0	0
OSBPL11	0	34	0	0	0	0	0	0
FAM162A	0	33	0	0	0	0	0	0
DNAJB8	0	32	0	0	0	0	0	0
ISY1	0	32	0	0	0	0	0	0
NUDT16	0	32	0	0	1	0	0	0
NDUFB4	0	30	0	0	1	0	0	0
QTRTD1	0	29	0	0	0	0	0	0
C3orf22	0	28	0	0	0	0	0	0
COX17	0	28	0	0	1	0	0	0
CPNE4	0	27	0	1	6	0	0	0

Gene symbol	Number of highly correlated pairs involved in		Number of SNPs					
			Aorta database			CATHGEN database		
	Aorta dataset	GDS1597 dataset	*P* value < 0.05	*P* value < 0.1	Total SNPs	*P* value < 0.05	*P* value < 0.1	Total SNPs
PIK3R4	0	26	0	3	5	0	0	0
SLC12A8	0	26	0	0	2	0	0	0
MUC13	0	25	0	1	1	0	0	0
TMEM108	0	25	0	1	1	0	0	0
UMPS	0	23	0	0	1	0	0	0
ZXDC	0	22	0	0	0	0	0	0
KIAA1407	0	20	0	0	1	0	0	0
CHCHD6	0	17	0	0	2	0	0	0
TMEM39A	0	17	0	0	1	0	0	0
RABL3	0	16	0	0	0	0	0	0
B4GALT4	0	12	0	0	1	0	0	0
ZDHHC23	0	12	2	4	11	0	0	0
ASTE1	0	11	0	0	0	0	0	0
MCM2	0	11	0	0	0	0	0	0
GPR175	0	9	0	0	0	0	0	0
POPDC2	0	9	0	0	0	0	0	0
NPHP3	0	8	0	0	1	0	0	0
PLA1A	0	8	2	2	5	0	0	0
GPR156	0	7	0	0	0	0	0	0

Gene symbol	Number of highly correlated pairs involved in		Number of SNPs					
			Aorta database			CATHGEN database		
	Aorta dataset	GDS1597 dataset	*P* value < 0.05	*P* value < 0.1	Total SNPs	*P* value < 0.05	*P* value < 0.1	Total SNPs
PLXND1	0	7	0	2	17	0	0	0
UROC1	0	7	0	0	0	0	0	0
C3orf1	0	6	0	0	2	0	0	0
RAB43	0	6	0	0	0	0	0	0
ADPRH	0	5	0	0	2	0	0	0
WDR5B	0	5	0	0	1	0	0	0
ABTB1	0	4	0	0	0	0	0	0
CCRL1	0	4	0	0	0	0	0	0
NEK11	0	4	1	1	3	0	0	0
PARP15	0	4	0	0	1	0	0	0
SEMA5B	0	4	0	0	1	0	0	0
KPNA1	0	2	0	0	1	0	0	0
SEC22A	0	2	0	0	0	0	0	0
UBA5	0	2	0	0	0	0	0	0
COL29A1	0	1	0	0	0	0	0	0
DTX3L	0	1	0	0	0	0	0	0
EEFSEC	0	1	0	0	0	0	0	0
GRAMD1C	0	1	0	0	0	0	0	0
IGSF11	0	1	4	4	11	0	0	0

Gene symbol	Number of highly correlated pairs involved in		Number of SNPs					
			Aorta database			CATHGEN database		
	Aorta dataset	GDS1597 dataset	*P* value < 0.05	*P* value < 0.1	Total SNPs	*P* value < 0.05	*P* value < 0.1	Total SNPs
KALRN	0	1	4	15	53	0	0	0

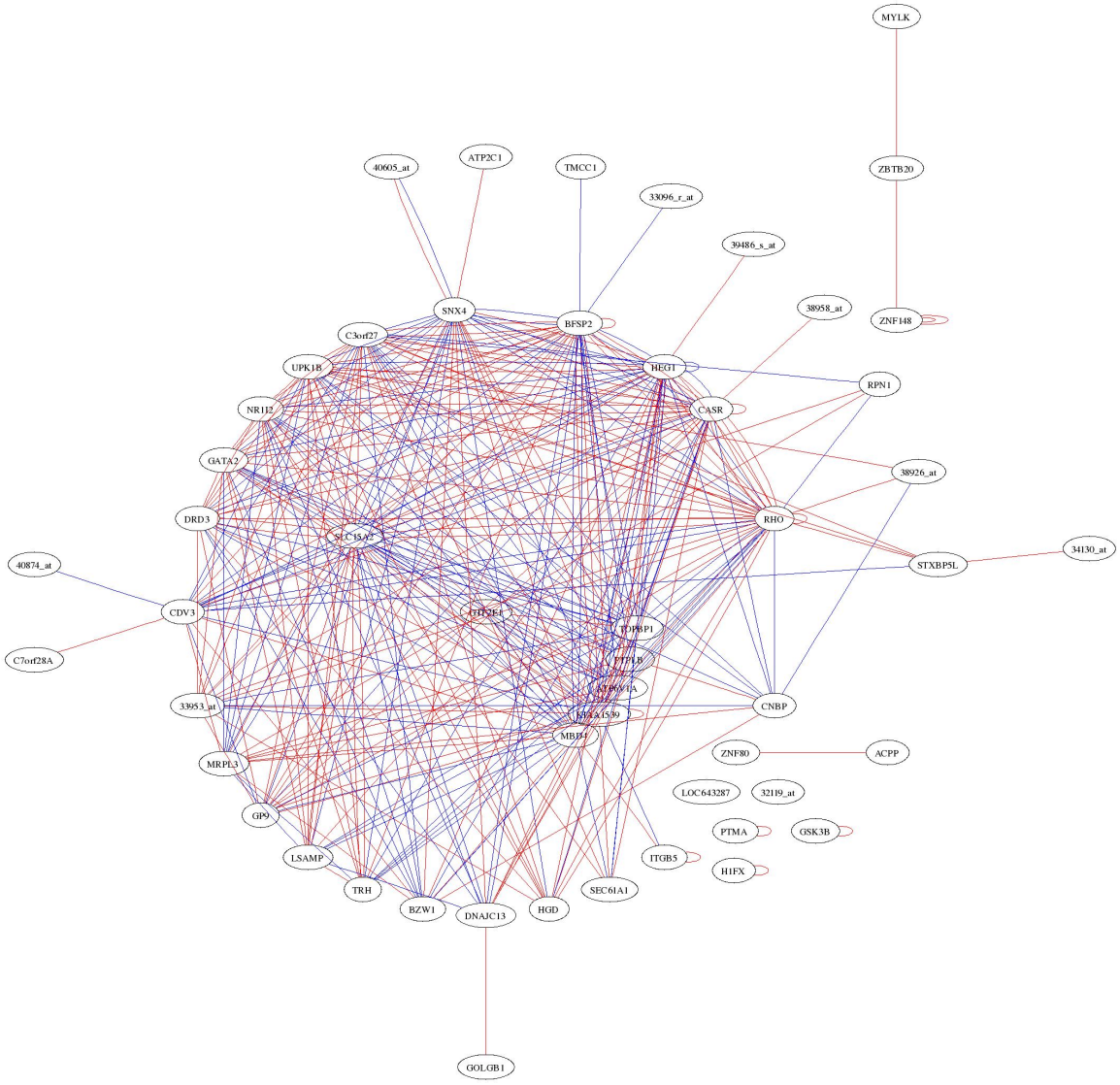
Appendix B

This table lists the genes involved in highly correlated pairs in the analysis result from GSE13985 dataset. The bolded genes represent overlaps in the gene lists from the Aorta dataset or GDS1597 dataset. The colored genes represent verified by DUMC SNPs data. Originally the list contains 259 tags. After removing 97 tags which are not corresponding to any genes, 162 genes were retained in the list.

Genes retained from GSE13985 dataset analysis						
ABTB1	CCDC37	FBXO40	KIAA1257	**MYLK**	**RHO**	**WDR5B**
ACAD11	CCDC48	FLJ40473	KIAA1333	NCRNA00119	ROPN1	**ZBTB20**
ACAD9	CCDC58	FSTL1	**KIAA1407**	**NEK11**	RPL32P3	**ZDHHC23**
ACPP	CD80	GAP43	KLF15	**NPHP3**	**RPN1**	**ZNF148**
ADCY5	CD86	**GATA2**	KLHDC6	**NR1I2**	RUVBL1	**ZNF80**
ADPRH	CDGAP	**GOLGB1**	**KPNA1**	**NUDT16**	**SEC22A**	**ZXDC**
ALDH1L1	**CDV3**	**GP9**	KTELC1	NUDT16P	**SEC61A1**	
AP1G2	**CHCHD6**	**GPR156**	LOC100128175	**OSBPL11**	**SEMA5B**	
ASTE1	CHST13	**GPR175**	LOC100129550	PARP14	**SLC12A8**	
ATP2C1	**CNBP**	**GRAMD1C**	LOC100129701	**PARP15**	**SLC15A2**	
B4GALT4	**COL29A1**	**GSK3B**	LOC100131117	PARP9	SLC25A23	
BFSP2	COL6A6	GSTO1	LOC100132207	PDIA5	SLC41A3	
BZW1	COPG	**GTF2E1**	LOC100132409	**PIK3R4**	**SNX4**	
C3orf1	**COX17**	H1FOO	LOC254057	**PLA1A**	**STXBP5L**	
C3orf15	**CPNE4**	**H1FX**	LOC285194	**PLXNA1**	TF	
C3orf22	CSRP2	HCLS1	LOC285224	PLXNA3	TIGIT	
C3orf25	CSTA	**HEG1**	LOC339874	**PLXND1**	**TMCC1**	
C3orf27	DIRC2	**HGD**	LOC90246	PMS2	**TMEM108**	
C3orf30	**DNAJB8**	HSPBAP1	LRRC58	PODXL2	**TMEM39A**	
C3orf37	**DNAJC13**	IFT122	**LSAMP**	POLQ	**TOPBP1**	
C3orf46	**DRD3**	**IGSF11**	**MBD4**	**POPDC2**	**TRH**	
C3orf47	**DTX3L**	ILDR1	**MCM2**	**PTPLB**	TXNRD3	
C3orf56	EAF2	IQCB1	MGC2848	**QTRTD1**	**UBA5**	
C7orf28A	**EEFSEC**	**ISY1**	MGLL	**RAB43**	**UMPS**	
CASR	EIF4E2	**ITGB5**	**MRPL3**	RAB7A	**UPK1B**	
CCDC14	**FAM162A**	**KALRN**	**MUC13**	**RABL3**	**UROC1**	

Appendix C

The figure illustrates that highly correlated pairs in the region of 115Mb to 135Mb on chromosome 3 in the Aorta dataset. The red edges mean positive correlations, and the blue edges mean negative correlations. We can discover that most of the correlations follow the rule of "*cis*-direction".



References

[1] Gibbons, G. H., & Dzau, V. J. (1996). Molecular therapies for vascular diseases. *Science, 272*(5262), 689-693.

[2] Wang, L., Hauser, E. R., Shah, S. H., Pericak-Vance, M. A., Haynes, C., Crosslin, D., Harris, M., Nelson, S., Hale, A. B., Granger, C. B., Haines, J. L., Jones, C. J., Crossman, D., Seo, D., Gregory, S. G., Kraus, W. E., Goldschmidt-Clermont, P. J., & Vance, J. M. (2007). Peakwide mapping on chromosome 3q13 identifies the kalirin gene as a novel candidate gene for coronary artery disease. *Am J Hum Genet, 80*(4), 650-663.

[3] Licatalosi, D. D., & Darnell, R. B. (2010). RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet, 11*(1), 75-87.

[4] Tuteja, N., & Tuteja, R. (2001). Unraveling DNA repair in human: molecular mechanisms and consequences of repair defect. *Crit Rev Biochem Mol Biol, 36*(3), 261-290.

[5] Roulston, A., Marcellus, R. C., & Branton, P. E. (1999). Viruses and apoptosis. *Annu Rev Microbiol, 53*, 577-628.

[6] Apostolakis, S., Amanatidou, V., Papadakis, E. G., & Spandidos, D. A. (2009). Genetic diversity of CX3CR1 gene and coronary artery disease: new insights through a meta-analysis. *Atherosclerosis, 207*(1), 8-15.

[7] Benn, M. (2009). Apolipoprotein B levels, APOB alleles, and risk of ischemic cardiovascular disease in the general population, a review. *Atherosclerosis, 206*(1), 17-30.

[8] Ding, K., & Kullo, I. J. (2009). Genome-wide association studies for atherosclerotic vascular disease and its risk factors. *Circ Cardiovasc Genet, 2*(1), 63-72.

[9] Hamsten, A., & Eriksson, P. (2008). Identifying the susceptibility genes for coronary artery disease: from hyperbole through doubt to cautious optimism. *J Intern Med, 263*(5), 538-552.

[10] Iatan, I., Alrasadi, K., Ruel, I., Alwaili, K., & Genest, J. (2008). Effect of ABCA1 mutations on risk for myocardial infarction. *Curr Atheroscler Rep, 10*(5), 413-426.

[11] Cook, E. H., Jr., & Scherer, S. W. (2008). Copy-number variations associated with neuropsychiatric conditions. *Nature, 455*(7215), 919-923.

[12] Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., Hurles, M. E., Carter, N. P., Scherer, S. W., & Lee, C. (2006). Copy number variation: new insights in genome diversity. *Genome Res, 16*(8), 949-961.

[13] Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., & Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature, 444*(7118), 444-454.

[14] Abboud, S., Karhunen, P. J., Lutjohann, D., Goebeler, S., Luoto, T., Friedrichs, S., Lehtimaki, T., Pandolfo, M., & Laaksonen, R. (2007). Proprotein convertase subtilisin/kexin type 9 (PCSK9) gene is a risk factor of large-vessel atherosclerosis stroke. *PloS One, 2*(10), e1043.

[15] Armstrong, C., Abilleira, S., Sitzer, M., Markus, H. S., & Bevan, S. (2007). Polymorphisms in MMP family and TIMP genes and carotid artery intima-media thickness. *Stroke, 38*(11), 2895-2899.

[16] Kamstrup, P. R. (2010). Lipoprotein(a) and ischemic heart disease-A causal association? A review. *Atherosclerosis*.

[17] Lanktree, M., & Hegele, R. A. (2008). Copy number variation in metabolic phenotypes. *Cytogenet Genome Res, 123*(1-4), 169-175.

[18] Mihalich, A., Magnaghi, P., Sessa, L., Trubia, M., Acquati, F., & Taramelli, R. (1997). Genomic structure and organization of kringles type 3 to 10 of the apolipoprotein(a) gene in 6q26-27. *Gene, 196*(1-2), 1-8.

[19] Sun, Y. V., Peyser, P. A., & Kardia, S. L. (2009). A common copy number variation on chromosome 6 association with the gene expression level of

endothelin 1 in transformed B lymphocytes from three racial groups. *Circ Cardiovasc Genet, 2*(5), 483-488.

[20] Seo, D., Wang, T., Dressman, H., Herderick, E. E., Iversen, E. S., Dong, C., Vata, K., Milano, C. A., Rigat, F., Pittman, J., Nevins, J. R., West, M., & Goldschmidt-Clermont, P. J. (2004). Gene expression phenotypes of atherosclerosis. *Arterioscler Thromb Vasc Biol, 24*(10), 1922-1927.

[21] Glazier, A. M., Nadeau, J. H., & Aitman, T. J. (2002). Finding genes that underlie complex traits. *Science, 298*(5602), 2345-2349.

[22] Guttmacher, A. E., & Collins, F. S. (2002). Genomic medicine--a primer. *N Engl J Med, 347*(19), 1512-1520.

[23] Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet, 6*(2), 95-108.

[24] Wang, W. Y., Barratt, B. J., Clayton, D. G., & Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet, 6*(2), 109-118.

[25] Ge, X., Yamamoto, S., Tsutsumi, S., Midorikawa, Y., Ihara, S., Wang, S. M., & Aburatani, H. (2005). Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics, 86*(2), 127-141.

[26] Shyamsundar, R., Kim, Y. H., Higgins, J. P., Montgomery, K., Jorden, M., Sethuraman, A., van de Rijn, M., Botstein, D., Brown, P. O., & Pollack, J. R. (2005). A DNA microarray survey of gene expression in normal human tissues. *Genome Biol, 6*(3), R22.

[27] Yoon, S. S., Segal, N. H., Park, P. J., Detwiller, K. Y., Fernando, N. T., Ryeom, S. W., Brennan, M. F., & Singer, S. (2006). Angiogenic profile of soft tissue sarcomas based on analysis of circulating factors and microarray gene expression. *J Surg Res, 135*(2), 282-290.

[28] Hauser, E. R., Crossman, D. C., Granger, C. B., Haines, J. L., Jones, C. J., Mooser, V., McAdam, B., Winkelmann, B. R., Wiseman, A. H., Muhlestein, J. B., Bartel, A. G., Dennis, C. A., Dowdy, E., Estabrooks, S., Eggleston, K., Francis, S., Roche, K., Clevenger, P. W., Huang, L., Pedersen, B., Shah, S., Schmidt, S., Haynes, C., West, S., Asper, D., Booze, M., Sharma, S., Sundseth, S., Middleton, L., Roses, A. D.,

Hauser, M. A., Vance, J. M., Pericak-Vance, M. A., & Kraus, W. E. (2004). A genomewide scan for early-onset coronary artery disease in 438 families: the GENECARD Study. *Am J Hum Genet, 75*(3), 436-447.

[29] Connelly, J. J., Wang, T., Cox, J. E., Haynes, C., Wang, L., Shah, S. H., Crosslin, D. R., Hale, A. B., Nelson, S., Crossman, D. C., Granger, C. B., Haines, J. L., Jones, C. J., Vance, J. M., Goldschmidt-Clermont, P. J., Kraus, W. E., Hauser, E. R., & Gregory, S. G. (2006). GATA2 is associated with familial early-onset coronary artery disease. *PLoS Genet, 2*(8), e139.

[30] Horne, B. D., Hauser, E. R., Wang, L., Muhlestein, J. B., Anderson, J. L., Carlquist, J. F., Shah, S. H., & Kraus, W. E. (2009). Validation study of genetic associations with coronary artery disease on chromosome 3q13-21 and potential effect modification by smoking. *Ann Hum Genet, 73*(Pt 6), 551-558.

[31] Wang, L., Hauser, E. R., Shah, S. H., Seo, D., Sivashanmugam, P., Exum, S. T., Gregory, S. G., Granger, C. B., Haines, J. L., Jones, C. J., Crossman, D., Haynes, C., Kraus, W. E., Freedman, N. J., Pericak-Vance, M. A., Goldschmidt-Clermont, P. J., & Vance, J. M. (2008). Polymorphisms of the tumor suppressor gene LSAMP are associated with left main coronary artery disease. *Ann Hum Genet, 72*(Pt 4), 443-453.

[32] Roberts, R. (2008). Genetics of premature myocardial infarction. *Curr Atheroscler Rep, 10*(3), 186-193.

[33] Broeckel, U., Hengstenberg, C., Mayer, B., Holmer, S., Martin, L. J., Comuzzie, A. G., Blangero, J., Nurnberg, P., Reis, A., Riegger, G. A., Jacob, H. J., & Schunkert, H. (2002). A comprehensive linkage analysis for myocardial infarction and its related risk factors. *Nat Genet, 30*(2), 210-214.

[34] Francke, S., Manraj, M., Lacquemant, C., Lecoeur, C., Lepretre, F., Passa, P., Hebe, A., Corset, L., Yan, S. L., Lahmidi, S., Jankee, S., Gunness, T. K., Ramjuttun, U. S., Balgobin, V., Dina, C., & Froguel, P. (2001). A genome-wide scan for coronary heart disease suggests in Indo-Mauritians a susceptibility locus on chromosome 16p13 and replicates linkage with the metabolic syndrome on 3q27. *Hum Mol Genet, 10*(24), 2751-2765.

[35] Harrap, S. B., Zammit, K. S., Wong, Z. Y., Williams, F. M., Bahlo, M., Tonkin, A. M., & Anderson, S. T. (2002). Genome-wide linkage analysis of the acute

coronary syndrome suggests a locus on chromosome 2. *Arterioscler Thromb Vasc Biol, 22*(5), 874-878.

[36] Shalon, D., Smith, S. J., & Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res, 6*(7), 639-645.

[37] Dettling, M., Gabrielson, E., & Parmigiani, G. (2005). Searching for differentially expressed gene combinations. *Genome Biol, 6*(10), R88.

[38] Goeman, J. J., van de Geer, S. A., de Kort, F., & van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics, 20*(1), 93-99.

[39] Jaeger, J., Sengupta, R., & Ruzzo, W. L. (2003). Improved gene selection for classification of microarrays. *Pac Symp Biocomput*, 53-64.

[40] Klebanov, L., Jordan, C., & Yakovlev, A. (2006). A new type of stochastic dependence revealed in gene expression data. *Stat Appl Genet Mol Biol, 5*, Article7.

[41] Lu, Y., Liu, P. Y., Xiao, P., & Deng, H. W. (2005). Hotelling's T2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics, 21*(14), 3105-3113.

[42] Relationship of atherosclerosis in young men to serum lipoprotein cholesterol concentrations and smoking. A preliminary report from the Pathobiological Determinants of Atherosclerosis in Youth (PDAY) Research Group. (1990). *JAMA, 264*(23), 3018-3024.

[43] Fortin, D. F., Califf, R. M., Pryor, D. B., & Mark, D. B. (1995). The way of the future redux. *Am J Cardiol, 76*(16), 1177-1182.

[44] King, J. Y., Ferrara, R., Tabibiazar, R., Spin, J. M., Chen, M. M., Kuchinsky, A., Vailaya, A., Kincaid, R., Tsalenko, A., Deng, D. X., Connolly, A., Zhang, P., Yang, E., Watt, C., Yakhini, Z., Ben-Dor, A., Adler, A., Bruhn, L., Tsao, P., Quertermous, T., & Ashley, E. A. (2005). Pathway analysis of coronary atherosclerosis. *Physiol Genomics, 23*(1), 103-118.

[45] Beisvag, V., Lehre, P. K., Midelfart, H., Aass, H., Geiran, O., Sandvik, A. K., Laegreid, A., Komorowski, J., & Ellingsen, O. (2006). Aetiology-specific patterns in end-stage heart failure patients identified by functional annotation and classification of microarray data. *Eur J Heart Fail, 8*(4), 381-389.

[46] Rader, D. J., Cohen, J., & Hobbs, H. H. (2003). Monogenic hypercholesterolemia: new insights in pathogenesis and treatment. *J Clin Invest, 111*(12), 1795-1803.

[47] Park, J. J., Berggren, J. R., Hulver, M. W., Houmard, J. A., & Hoffman, E. P. (2006). GRB14, GPD1, and GDF8 as potential network collaborators in weight loss-induced improvements in insulin action in human skeletal muscle. *Physiol Genomics, 27*(2), 114-121.

[48] West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Jr., Marks, J. R., & Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A, 98*(20), 11462-11467.

[49] Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., & Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behav Brain Res, 125*(1-2), 279-284.

[50] Reiner, A., Yekutieli, D., & Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics, 19*(3), 368-375.

[51] Sutton, B. S., Crosslin, D. R., Shah, S. H., Nelson, S. C., Bassil, A., Hale, A. B., Haynes, C., Goldschmidt-Clermont, P. J., Vance, J. M., Seo, D., Kraus, W. E., Gregory, S. G., & Hauser, E. R. (2008). Comprehensive genetic analysis of the platelet activating factor acetylhydrolase (PLA2G7) gene and cardiovascular disease in case-control and family datasets. *Hum Mol Genet, 17*(9), 1318-1328.

[52] Eagleson, K. L., Pimenta, A. F., Burns, M. M., Fairfull, L. D., Cornuet, P. K., Zhang, L., & Levitt, P. (2003). Distinct domains of the limbic system-associated membrane protein (LAMP) mediate discrete effects on neurite outgrowth. *Mol Cell Neurosci, 24*(3), 725-740.

[53] McNamee, C. J., Reed, J. E., Howard, M. R., Lodge, A. P., & Moss, D. J. (2002). Promotion of neuronal cell adhesion by members of the IgLON family occurs in

the absence of either support or modification of neurite outgrowth. *J Neurochem, 80*(6), 941-948.

[54] Pimenta, A. F., Zhukareva, V., Barbe, M. F., Reinoso, B. S., Grimley, C., Henzel, W., Fischer, I., & Levitt, P. (1995). The limbic system-associated membrane protein is an Ig superfamily member that mediates selective neuronal growth and axon targeting. *Neuron, 15*(2), 287-297.

[55] Chen, J., Lui, W. O., Vos, M. D., Clark, G. J., Takahashi, M., Schoumans, J., Khoo, S. K., Petillo, D., Lavery, T., Sugimura, J., Astuti, D., Zhang, C., Kagawa, S., Maher, E. R., Larsson, C., Alberts, A. S., Kanayama, H. O., & Teh, B. T. (2003). The t(1;3) breakpoint-spanning genes LSAMP and NORE1 are involved in clear cell renal cell carcinomas. *Cancer Cell, 4*(5), 405-413.

[56] Ntougkos, E., Rush, R., Scott, D., Frankenberg, T., Gabra, H., Smyth, J. F., & Sellar, G. C. (2005). The IgLON family in epithelial ovarian cancer: expression profiles and clinicopathologic correlates. *Clin Cancer Res, 11*(16), 5764-5768.

[57] Hu, Z., Bonifas, J. M., Beech, J., Bench, G., Shigihara, T., Ogawa, H., Ikeda, S., Mauro, T., & Epstein, E. H., Jr. (2000). Mutations in ATP2C1, encoding a calcium pump, cause Hailey-Hailey disease. *Nat Genet, 24*(1), 61-65.

[58] Okunade, G. W., Miller, M. L., Azhar, M., Andringa, A., Sanford, L. P., Doetschman, T., Prasad, V., & Shull, G. E. (2007). Loss of the Atp2c1 secretory pathway Ca(2+)-ATPase (SPCA1) in mice causes Golgi stress, apoptosis, and midgestational death in homozygous embryos and squamous cell tumors in adult heterozygotes. *J Biol Chem, 282*(36), 26517-26527.

[59] Linstedt, A. D., & Hauri, H. P. (1993). Giantin, a novel conserved Golgi membrane protein containing a cytoplasmic domain of at least 350 kDa. *Mol Biol Cell, 4*(7), 679-693.

[60] Brown, E. M., Gamba, G., Riccardi, D., Lombardi, M., Butters, R., Kifor, O., Sun, A., Hediger, M. A., Lytton, J., & Hebert, S. C. (1993). Cloning and characterization of an extracellular Ca(2+)-sensing receptor from bovine parathyroid. *Nature, 366*(6455), 575-580.

[61] Molostvov, G., Fletcher, S., Bland, R., & Zehnder, D. (2008). Extracellular calcium-sensing receptor mediated signalling is involved in human vascular

smooth muscle cell proliferation and apoptosis. *Cell Physiol Biochem, 22*(5-6), 413-422.

[62] Motoyama, H. I., & Friedman, P. A. (2002). Calcium-sensing receptor regulation of PTH-dependent calcium absorption by mouse cortical ascending limbs. *Am J Physiol Renal Physiol, 283*(3), F399-406.

[63] Romoli, R., Lania, A., Mantovani, G., Corbetta, S., Persani, L., & Spada, A. (1999). Expression of calcium-sensing receptor and characterization of intracellular signaling in human pituitary adenomas. *J Clin Endocrinol Metab, 84*(8), 2848-2853.

[64] Brown, E. M., Vassilev, P. M., Quinn, S., & Hebert, S. C. (1999). G-protein-coupled, extracellular Ca(2+)-sensing receptor: a versatile regulator of diverse cellular functions. *Vitam Horm, 55*, 1-71.

[65] Olszak, I. T., Poznansky, M. C., Evans, R. H., Olson, D., Kos, C., Pollak, M. R., Brown, E. M., & Scadden, D. T. (2000). Extracellular calcium elicits a chemokinetic response from monocytes in vitro and in vivo. *J Clin Invest, 105*(9), 1299-1305.

[66] Tfelt-Hansen, J., & Brown, E. M. (2005). The calcium-sensing receptor in normal physiology and pathophysiology: a review. *Crit Rev Clin Lab Sci, 42*(1), 35-70.

[67] Fukushige, S., Kondo, E., Gu, Z., Suzuki, H., & Horii, A. (2006). RET finger protein enhances MBD2- and MBD4-dependent transcriptional repression. *Biochem Biophys Res Commun, 351*(1), 85-92.

[68] Hendrich, B., Hardeland, U., Ng, H. H., Jiricny, J., & Bird, A. (1999). The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature, 401*(6750), 301-304.

[69] Kondo, E., Gu, Z., Horii, A., & Fukushige, S. (2005). The thymine DNA glycosylase MBD4 represses transcription and is associated with methylated p16(INK4a) and hMLH1 genes. *Mol Cell Biol, 25*(11), 4388-4396.

[70] Miao, R., Gu, H., Liu, H., Hu, Z., Jin, G., Wang, H., Wang, Y., Sun, W., Ma, H., Chen, D., Tian, T., Jin, L., Wei, Q., Lu, D., Huang, W., & Shen, H. (2008). Tagging single nucleotide polymorphisms in MBD4 are associated with risk of lung cancer in a Chinese population. *Lung Cancer, 62*(3), 281-286.

[71] Song, J. H., Maeng, E. J., Cao, Z., Kim, S. Y., Nam, S. W., Lee, J. Y., & Park, W. S. (2009). The Glu346Lys polymorphism and frameshift mutations of the Methyl-CpG Binding Domain 4 gene in gastrointestinal cancer. *Neoplasma, 56*(4), 343-347.

[72] Andreassi, M. G. (2003). Coronary atherosclerosis and somatic mutations: an overview of the contributive factors for oxidative DNA damage. *Mutat Res, 543*(1), 67-86.

[73] De Flora, S., Izzotti, A., Walsh, D., Degan, P., Petrilli, G. L., & Lewtas, J. (1997). Molecular epidemiology of atherosclerosis. *FASEB J, 11*(12), 1021-1031.

[74] Evans, M. D., Dizdaroglu, M., & Cooke, M. S. (2004). Oxidative DNA damage and disease: induction, repair and significance. *Mutat Res, 567*(1), 1-61.

[75] Gackowski, D., Kruszewski, M., Jawien, A., Ciecierski, M., & Olinski, R. (2001). Further evidence that oxidative stress may be a risk factor responsible for the development of atherosclerosis. *Free Radic Biol Med, 31*(4), 542-547.

[76] Izzotti, A., De Flora, S., Petrilli, G. L., Gallagher, J., Rojas, M., Alexandrov, K., Bartsch, H., & Lewtas, J. (1995). Cancer biomarkers in human atherosclerotic lesions: detection of DNA adducts. *Cancer Epidemiol Biomarkers Prev, 4*(2), 105-110.

[77] Van Schooten, F. J., Hirvonen, A., Maas, L. M., De Mol, B. A., Kleinjans, J. C., Bell, D. A., & Durrer, J. D. (1998). Putative susceptibility markers of coronary artery disease: association between VDR genotype, smoking, and aromatic DNA adduct levels in human right atrial tissue. *FASEB J, 12*(13), 1409-1417.

[78] Yoon, J. H., Iwai, S., O'Connor, T. R., & Pfeifer, G. P. (2003). Human thymine DNA glycosylase (TDG) and methyl-CpG-binding protein 4 (MBD4) excise thymine glycol (Tg) from a Tg:G mispair. *Nucleic Acids Res, 31*(18), 5399-5404.

[79] Carter, C. J. (2007). Convergence of genes implicated in Alzheimer's disease on the cerebral cholesterol shuttle: APP, cholesterol, lipoproteins, and atherosclerosis. *Neurochem Int, 50*(1), 12-38.

[80] Zhu, W., Yang, L., & Du, Z. (2009). Layered functional network analysis of gene expression in human heart failure. *PloS One, 4*(7), e6288.

[81] Kerkela, R., Kockeritz, L., Macaulay, K., Zhou, J., Doble, B. W., Beahm, C., Greytak, S., Woulfe, K., Trivedi, C. M., Woodgett, J. R., Epstein, J. A., Force, T., & Huggins, G. S. (2008). Deletion of GSK-3beta in mice leads to hypertrophic cardiomyopathy secondary to cardiomyoblast hyperproliferation. *J Clin Invest, 118*(11), 3609-3618.

[82] Briegel, K., Lim, K. C., Plank, C., Beug, H., Engel, J. D., & Zenke, M. (1993). Ectopic expression of a conditional GATA-2/estrogen receptor chimera arrests erythroid differentiation in a hormone-dependent manner. *Genes Dev, 7*(6), 1097-1109.

[83] Tsai, F. Y., Keller, G., Kuo, F. C., Weiss, M., Chen, J., Rosenblatt, M., Alt, F. W., & Orkin, S. H. (1994). An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature, 371*(6494), 221-226.

[84] Tsai, F. Y., & Orkin, S. H. (1997). Transcription factor GATA-2 is required for proliferation/survival of early hematopoietic cells and mast cell formation, but not for erythroid and myeloid terminal differentiation. *Blood, 89*(10), 3636-3643.

[85] Cernuda-Morollon, E., & Ridley, A. J. (2006). Rho GTPases and leukocyte adhesion receptor expression and function in endothelial cells. *Circ Res, 98*(6), 757-767.

[86] Knaus, U. G. (2000). Rho GTPase signaling in inflammation and transformation. *Immunol Res, 21*(2-3), 103-109.

[87] Schwartz, M. (2004). Rho signalling at a glance. *J Cell Sci, 117*(Pt 23), 5457-5458.

[88] Wherlock, M., & Mellor, H. (2002). The Rho GTPase family: a Racs to Wrchs story. *J Cell Sci, 115*(Pt 2), 239-240.

[89] Noma, K., Oyama, N., & Liao, J. K. (2006). Physiological role of ROCKs in the cardiovascular system. *Am J Physiol Cell Physiol, 290*(3), C661-668.

[90] Ren, J., & Fang, C. X. (2005). Small guanine nucleotide-binding protein Rho and myocardial function. *Acta Pharmacol Sin, 26*(3), 279-285.

[91] Shimokawa, H., & Takeshita, A. (2005). Rho-kinase is an important therapeutic target in cardiovascular medicine. *Arterioscler Thromb Vasc Biol, 25*(9), 1767-1775.

[92] Tzima, E. (2006). Role of small GTPases in endothelial cytoskeletal dynamics and the shear stress response. *Circ Res, 98*(2), 176-185.

[93] Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., & Lee, C. (2004). Detection of large-scale variation in the human genome. *Nat Genet, 36*(9), 949-951.

[94] Ahn, S. M., Kim, T. H., Lee, S., Kim, D., Ghang, H., Kim, D. S., Kim, B. C., Kim, S. Y., Kim, W. Y., Kim, C., Park, D., Lee, Y. S., Kim, S., Reja, R., Jho, S., Kim, C. G., Cha, J. Y., Kim, K. H., Lee, B., Bhak, J., & Kim, S. J. (2009). The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res, 19*(9), 1622-1629.

[95] Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara, E. C. M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo,

M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., & Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature, 456*(7218), 53-59.

[96] Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., Macarthur, D. G., Macdonald, J. R., Onyiah, I., Pang, A. W., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W., & Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature, 464*(7289), 704-712.

[97] Cooper, G. M., Zerr, T., Kidd, J. M., Eichler, E. E., & Nickerson, D. A. (2008). Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet, 40*(10), 1199-1203.

[98] de Smith, A. J., Tsalenko, A., Sampas, N., Scheffer, A., Yamada, N. A., Tsang, P., Ben-Dor, A., Yakhini, Z., Ellis, R. J., Bruhn, L., Laderman, S., Froguel, P., & Blakemore, A. I. (2007). Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum Mol Genet, 16*(23), 2783-2794.

[99] Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M., & Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Res, 19*(2), 318-326.

[100] Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H. C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N. A., & Singleton, A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature, 451*(7181), 998-1003.

[101] Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tuzun, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R., & Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature, 453*(7191), 56-64.

[102] Kim, J. I., Ju, Y. S., Park, H., Kim, S., Lee, S., Yi, J. H., Mudge, J., Miller, N. A., Hong, D., Bell, C. J., Kim, H. S., Chung, I. S., Lee, W. C., Lee, J. S., Seo, S. H., Yun, J. Y., Woo, H. N., Lee, H., Suh, D., Kim, H. J., Yavartanoo, M., Kwak, M., Zheng, Y., Lee, M. K., Kim, J. Y., Gokcumen, O., Mills, R. E., Zaranek, A. W., Thakuria, J., Wu, X., Kim, R. W., Huntley, J. J., Luo, S., Schroth, G. P., Wu, T. D., Kim, H., Yang, K. S., Park, W. Y., Church, G. M., Lee, C., Kingsmore, S. F., & Seo, J. S. (2009). A highly annotated whole-genome sequence of a Korean individual. *Nature, 460*(7258), 1011-1015.

[103] Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, A. C., Chi, J., Yang, F., Carter, N. P., Hurles, M. E., Weissman, S. M., Harkins, T. T., Gerstein, M. B., Egholm, M., & Snyder, M. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science, 318*(5849), 420-426.

[104] Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., Busam, D. A., Beeson, K. Y., McIntosh, T. C., Remington, K. A., Abril, J. F., Gill, J., Borman, J., Rogers, Y. H., Frazier, M. E., Scherer, S. W., Strausberg, R. L., & Venter, J. C. (2007). The diploid genome sequence of an individual human. *PLoS Biol, 5*(10), e254.

[105] Matsuzaki, H., Wang, P. H., Hu, J., Rava, R., & Fu, G. K. (2009). High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians. *Genome Biol, 10*(11), R125.

[106] McCarroll, S. A., Hadnott, T. N., Perry, G. H., Sabeti, P. C., Zody, M. C., Barrett, J. C., Dallaire, S., Gabriel, S. B., Lee, C., Daly, M. J., & Altshuler, D. M. (2006). Common deletion polymorphisms in the human genome. *Nat Genet, 38*(1), 86-92.

[107] McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., de Bakker, P. I., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K. W., Rava, R., Daly, M. J., Gabriel, S. B., & Altshuler, D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet, 40*(10), 1166-1174.

[108] McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., Dimalanta, E. T., Hyland, F. C., Sokolsky, T. D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C. L., Li, B., Kotler, L., Stuart, J. R., Malek, J. A., Manning, J. M., Antipova, A. A., Perez, D. S., Moore, M. P., Hayashibara, K. C., Lyons, M. R., Beaudoin, R. E., Coleman, B. E., Laptewicz, M. W., Sannicandro, A. E., Rhodes, M. D., Gottimukkala, R. K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J. M., Eichler, E. E., Reese, M. G., De La Vega, F. M., & Blanchard, A. P. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res, 19*(9), 1527-1541.

[109] Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., & Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res, 16*(9), 1182-1190.

[110] Perry, G. H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C. W., Scheffer, A., Steinfeld, I., Tsang, P., Yamada, N. A., Park, H. S., Kim, J. I., Seo, J. S., Yakhini, Z., Laderman, S., Bruhn, L., & Lee, C. (2008). The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet, 82*(3), 685-695.

[111] Pinto, D., Marshall, C., Feuk, L., & Scherer, S. W. (2007). Copy-number variation in control population cohorts. *Hum Mol Genet, 16 Spec No. 2*, R168-173.

[112] Shaikh, T. H., Gai, X., Perin, J. C., Glessner, J. T., Xie, H., Murphy, K., O'Hara, R., Casalunovo, T., Conlin, L. K., D'Arcy, M., Frackelton, E. C., Geiger, E. A., Haldeman-Englert, C., Imielinski, M., Kim, C. E., Medne, L., Annaiah, K., Bradfield, J. P., Dabaghyan, E., Eckert, A., Onyiah, C. C., Ostapenko, S., Otieno, F.

G., Santa, E., Shaner, J. L., Skraban, R., Smith, R. M., Elia, J., Goldmuntz, E., Spinner, N. B., Zackai, E. H., Chiavacci, R. M., Grundmeier, R., Rappaport, E. F., Grant, S. F., White, P. S., & Hakonarson, H. (2009). High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res, 19*(9), 1682-1690.

[113] Simon-Sanchez, J., Scholz, S., Fung, H. C., Matarin, M., Hernandez, D., Gibbs, J. R., Britton, A., de Vrieze, F. W., Peckham, E., Gwinn-Hardy, K., Crawley, A., Keen, J. C., Nash, J., Borgaonkar, D., Hardy, J., & Singleton, A. (2007). Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. *Hum Mol Genet, 16*(1), 1-14.

[114] Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y. J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X. Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. A., & Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature, 452*(7189), 872-876.

[115] Wong, K. K., deLeeuw, R. J., Dosanjh, N. S., Kimm, L. R., Cheng, Z., Horsman, D. E., MacAulay, C., Ng, R. T., Brown, C. J., Eichler, E. E., & Lam, W. L. (2007). A comprehensive analysis of common copy-number variations in the human genome. *Am J Hum Genet, 80*(1), 91-104.

[116] Erdmann, J., Grosshennig, A., Braund, P. S., Konig, I. R., Hengstenberg, C., Hall, A. S., Linsel-Nitschke, P., Kathiresan, S., Wright, B., Tregouet, D. A., Cambien, F., Bruse, P., Aherrahrou, Z., Wagner, A. K., Stark, K., Schwartz, S. M., Salomaa, V., Elosua, R., Melander, O., Voight, B. F., O'Donnell, C. J., Peltonen, L., Siscovick, D. S., Altshuler, D., Merlini, P. A., Peyvandi, F., Bernardinelli, L., Ardissino, D., Schillert, A., Blankenberg, S., Zeller, T., Wild, P., Schwarz, D. F., Tiret, L., Perret, C., Schreiber, S., El Mokhtari, N. E., Schafer, A., Marz, W., Renner, W., Bugert, P., Kluter, H., Schrezenmeir, J., Rubin, D., Ball, S. G., Balmforth, A. J., Wichmann, H. E., Meitinger, T., Fischer, M., Meisinger, C., Baumert, J., Peters, A., Ouwehand, W. H., Deloukas, P., Thompson, J. R., Ziegler, A., Samani, N. J., & Schunkert, H. (2009). New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat Genet, 41*(3), 280-282.

[117] Banz, Y., Hess, O. M., Robson, S. C., Csizmadia, E., Mettler, D., Meier, P., Haeberli, A., Shaw, S., Smith, R. A., & Rieben, R. (2007). Attenuation of

myocardial reperfusion injury in pigs by Mirococept, a membrane-targeted complement inhibitor derived from human CR1. *Cardiovasc Res, 76*(3), 482-493.

[118] Boiocchi, C., Zorzetto, M., Sbarsi, I., Pirotta, A., Schirinzi, S., Falcone, C., & Cuccia, M. (2009). CR1 genotype and haplotype involvement in coronary artery disease: the pivotal role of hypertension and dyslipidemia. *Int J Mol Med, 24*(2), 181-187.

[119] Fairweather, D., Frisancho-Kiss, S., Njoku, D. B., Nyland, J. F., Kaya, Z., Yusung, S. A., Davis, S. E., Frisancho, J. A., Barrett, M. A., & Rose, N. R. (2006). Complement receptor 1 and 2 deficiency increases coxsackievirus B3-induced myocarditis, dilated cardiomyopathy, and heart failure by increasing macrophages, IL-1beta, and immune complex deposition in the heart. *J Immunol, 176*(6), 3516-3524.

[120] Karthikeyan, G., Baalasubramanian, S., Seth, S., & Das, N. (2007). Low levels of plasma soluble complement receptor type 1 in patients receiving thrombolytic therapy for acute myocardial infarction. *J Thromb Thrombolysis, 23*(2), 115-120.

[121] Sperr, W. R., Bankl, H. C., Mundigler, G., Klappacher, G., Grossschmidt, K., Agis, H., Simon, P., Laufer, P., Imhof, M., Radaszkiewicz, T., Glogar, D., Lechner, K., & Valent, P. (1994). The human cardiac mast cell: localization, isolation, phenotype, and functional characterization. *Blood, 84*(11), 3876-3884.

[122] Stahl, G. L., Reenstra, W. R., & Frendl, G. (1995). Complement-mediated loss of endothelium-dependent relaxation of porcine coronary arteries. Role of the terminal membrane attack complex. *Circ Res, 76*(4), 575-583.

[123] Vakeva, A., Laurila, P., & Meri, S. (1993). Regulation of complement membrane attack complex formation in myocardial infarction. *Am J Pathol, 143*(1), 65-75.

[124] Martin-Ventura, J. L., Madrigal-Matute, J., Munoz-Garcia, B., Blanco-Colio, L. M., Van Oostrom, M., Zalba, G., Fortuno, A., Gomez-Guerrero, C., Ortega, L., Ortiz, A., Diez, J., & Egido, J. (2009). Increased CD74 expression in human atherosclerotic plaques: contribution to inflammatory responses in vascular cells. *Cardiovasc Res, 83*(3), 586-594.

[125] Esposito, G., Grutter, G., Drago, F., Costa, M. W., De Santis, A., Bosco, G., Marino, B., Bellacchio, E., Lepri, F., Harvey, R. P., Sarkozy, A., & Dallapiccola, B. (2009). Molecular analysis of PRKAG2, LAMP2, and NKX2-5 genes in a cohort of

125 patients with accessory atrioventricular connection. *Am J Med Genet A, 149A*(7), 1574-1577.

[126] Harvey, R. P., Lai, D., Elliott, D., Biben, C., Solloway, M., Prall, O., Stennard, F., Schindeler, A., Groves, N., Lavulo, L., Hyun, C., Yeoh, T., Costa, M., Furtado, M., & Kirk, E. (2002). Homeodomain factor Nkx2-5 in heart development and disease. *Cold Spring Harb Symp Quant Biol, 67*, 107-114.

[127] Reamon-Buettner, S. M., & Borlak, J. (2004). Somatic NKX2-5 mutations as a novel mechanism of disease in complex congenital heart disease. *J Med Genet, 41*(9), 684-690.

[128] Toko, H., Zhu, W., Takimoto, E., Shiojima, I., Hiroi, Y., Zou, Y., Oka, T., Akazawa, H., Mizukami, M., Sakamoto, M., Terasaki, F., Kitaura, Y., Takano, H., Nagai, T., Nagai, R., & Komuro, I. (2002). Csx/Nkx2-5 is required for homeostasis and survival of cardiac myocytes in the adult heart. *J Biol Chem, 277*(27), 24735-24743.

[129] Turk, V., Stoka, V., & Turk, D. (2008). Cystatins: biochemical and structural properties, and medical relevance. *Front Biosci, 13*, 5406-5420.

[130] Flouris, G. A., Arvanitis, D. A., Parissis, J. T., Arvanitis, D. L., & Spandidos, D. A. (2000). Loss of heterozygosity in DNA mismatch repair genes in human atherosclerotic plaques. *Mol Cell Biol Res Commun, 4*(1), 62-65.

[131] Kardys, I., Uitterlinden, A. G., Hofman, A., Witteman, J. C., & de Maat, M. P. (2007). Fibrinogen gene haplotypes in relation to risk of coronary events and coronary and extracoronary atherosclerosis: the Rotterdam Study. *Thromb Haemost, 97*(2), 288-295.